ISSUES AND CHALLENGES IN VALIDATING MILITARY SIMULATION MODELS

THESIS

Michael R. Elmer
Captain, USAF
AFIT/GSO/ENS/95D-02

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GSO/ENS/95D-02

ISSUES AND CHALLENGES IN VALIDATING MILITARY SIMULATION MODELS

THESIS

Michael R. Elmer
Captain, USAF
AFIT/GSO/ENS/95D-02

THESIS APPROVAL

STUDENT: Capt Michael Elmer CLASS: GSO-95D

THESIS TITLE: ISSUES AND CHALLENGES IN VALIDATING MILITARY SIMULATION MODELS

DEFENSE DATE: 8 December 1995

COMMITTEE:	NAME/DEPARTMENT	SIGNATURE
Advisor	Lt Col Paul Auclair/ENS	Paul J Auclair
Advisor	Dr. Edward Mykytka/ENS	Edward J. Mykytka

AFIT/GSO/ENS/95D-02

ISSUES AND CHALLENGES IN VALIDATING MILITARY SIMULATION MODELS

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air Education and Training Command

In Partial Fulfillment of the

Requirements for the Degree of

Masters of Science in Space Operations

Michael R. Elmer, B.S.

Captain, USAF

December 1995

Approved for Public release; distribution unlimited

Preface

This study was started with the goal of examining the important issues of simulation validation that face military analysts. Because computer simulation use is becoming more prevalent in analysis in today's Air Force, proper validation of those models is important. There are many issues that military simulation analysts have to deal with, so this research is intended to sort through those issues and try to focus the effort of validation.

I would like to thank Lt Col Auclair and Dr. Mykitka for their support, patience, and direction as my co-advisors. I would like to send my sincerest gratitude to my parents Otto and Marie Elmer. For all the times they were there when I needed support, and all the years that they spent instilling in me the values and qualities that I have relied on to get me to this point, I thank you.

Michael Elmer

Table of Contents

List of Figures

List of Tables

Table

AFIT/GSO/ENS/95D-02

Abstract

The purpose of this thesis is to address the challenges of validating simulation models, especially those challenges facing military simulation analysts. Three distinct issues are of concern to the military simulation analyst: 1) What type of validation effort do the academic experts recommend? 2) What does the military policy say is necessary for a proper validation effort? 3) What can a simulation practitioner realistically accomplish given time and resource constraints?

Four methodologies were chosen to represent the academic perspective on validation. A model of validation methods is integrated from the methodologies of these four simulation validation references.

The validation policies of the Army, Navy, and Air Force are examined and analyzed for their methodologies to be applied to simulations. The integrated model is compared to these policies that are being formed within the DoD to determine the relationship between the academic experts and the military policies.

Case studies of validation efforts are examined and analyzed for the methodologies used by simulation practitioners. The integrated model is compared to the case studies to examine the relationship between the academic experts and the actual practitioners. Finally, conclusions and observations are drawn from all of these comparisons.

Focusing the Issues and Challenges of Military Simulation Validation

1. INTRODUCTION

The validation of simulation models has been an elusive art since the advent of simulation modeling. Many papers have been written on validation, but there are few actual case studies on the subject. (Kleijnen, 1995)

The classical definition of validation is the determination of whether or not the conceptual model used in a simulation is an accurate representation of the system under study. (Law and Kelton, 1991) This definition can be open to interpretation. What is meant by "an accurate representation"? A simulation model is an abstraction of a real world system, therefore it can never represent the real system exactly. Details of the real system that are not pertinent to the problem can be excluded. For example, consider the factors that affect the flight of an F-16. The moon exerts gravitational forces on the F-16, but they are negligible compared to those of the Earth and can be left out of a simulation. In contrast, a model of a satellite orbiting the Earth in a geosynchronous orbit should include the gravitational forces from the moon. Ignoring these forces could lead to a grossly inaccurate model and misleading results, which in turn could have a major impact on the decisions based on the simulation study.

A better definition of validation, proposed by Kleijnen, might be that validation is the process of determining if the conceptual model is 'good enough' for use, which depends on the goals of the simulation. (Kleijnen, 1995) The effectiveness of the

simulation validation process is related as a degree of confidence that the validating analyst has in the conceptual model being *'good enough' for use to achieve the goals of the simulation*. (Balci, 1994) However, the word *'degree'* implies that there a specific quantitative measure that can be applied to the model. None of the references examined in this thesis suggested a method of applying such a measure. To be clear on the terminology used, the definition of validation will be stated as follows; validation is the process of determining if a conceptual model is suitable for use to achieve the goals of the particular simulation.

The Air Force document addressing verification, validation, and accreditation (VV&A) of simulation models, *Air Force Instruction 16-1001*, which is in draft form at this writing, states that "validation is the rigorous and structured process of determining the extent to which models and simulations (M&S) accurately represent the real-world phenomenon from the prospective M&S use." The AF policy emphasizes that the validation effort includes examination of all algorithms, assumptions, and structure, in the context of the model's intended use.

Air Force Instruction 16-1001 requires that a documented validation effort is made on models that fit any of the following criteria:

1. Engagement, mission, or any campaign level models that will be briefed to senior ranking officials outside of the Air Force;
2. Models used significantly in a cost and operational effectiveness analysis;
3. Models used for force structure, resources, warfare requirements, and assessment analysis;
4. Models used in acquisition projects involving over $115 million in research, test, design, and evaluation or $540 million in procurement;

5. Models used for 'real time' control and movement of troops;
6. Models with aspects dealing with human safety;
7. Models made available to agencies outside the Air Force that Air Force Directorate of Modeling, Simulation, and Analysis (AF/XOM) has determined warrants the attention.

These criteria apply to many projects, but not all. For smaller simulation projects, the Instruction does not mandate validation, but rather leaves it to the decision of the major command that the project falls under. For projects that these criteria do apply to, the Instruction only suggests possible validation techniques. Here lies the main problem, there is no clear guidance on the type of validation effort that is required. A 4-star summit on modeling and simulation recently cited "no implemented verification, validation and accreditation process" as a quality deficiency Air Force wide.[1] A DoD Inspector General audit reported that 95% of all DoD models and simulations that had been inspected had not been validated in a structured manner. (Piplani, 1994, pg 6-3) One of the recommendations from the report was development of policy for standards for validation.

Verification is defined as the determination that a simulation performs as intended. (Law and Kelton, 1991) Verification is presented here because this effort is often times performed concurrently with validation, but it is distinctly different. Where validation is determining the aptness of the conceptual model for use, verification is the process of determining whether or not the conceptual model has been accurately implemented as a computer or mathematical model. (Law and Kelton, 1991) Verification of models will not be discussed in this thesis.

[1] U.S. Air Force 4-Star Modeling and Simulation Summit, 9 June 1995, Andrews AFB, MD.

When investigating validation, there are at least three clearly defined perspectives that the military analyst should consider. The first view is the 'idealized' world of academia. What kind of effort do academic experts say is needed? With no real constraints, experts in the academic world can develop virtually unlimited lists of procedures to follow. The second view is military policy. As already stated, the policy is currently being created. The final view is that of the practitioner. What can a simulation practitioner really do with the resources available and the time allotted? What are other analysts doing, if anything, to validate simulation models? A growing number of managers are interested in using simulation as an integral part of their decision processes. Significant decisions based on results from simulation models require verified and validated models. The challenge to the simulation developer and customer is to balance validation requirements against time, funds, and manpower constraints.

The purpose of this thesis is to address the challenges of validating simulation models, especially those challenges facing military simulation analysts. This thesis presents a model of validation steps that is an integration of views of simulation experts in the academic world. This integrated model is compared to the policies that are being formed within the DoD, and with validation efforts that were actually performed and published as case studies.

This thesis will progress in the following manner. Chapter 2 is a review and comparison of academic work on validation, resulting in the construction of an integrated, validation methodology model. Chapter 3 is a review of military policy and how it compares with the integrated academic model. Chapter 4 is an evaluation of case studies

of validation efforts and comparisons of the methodologies used in the case studies with the integrated model. Finally, Chapter 5 is a presentation of conclusions developed from this effort.

2. BACKGROUND

2.1 Introduction

Validation is a complex subject that requires the simulation analyst not only to have specific knowledge of validation, but also challenges the analyst to use insight and creativity. (Balci, 1994) In this chapter, four academic works in simulation validation are reviewed, and a validation methodology model is created that integrates the methodologies from the four academic works. This integrated model is used in following chapters in a comparison with military policy methodology and then with published case studies of validation efforts. From these comparisons, conclusions will be formed of what methodology steps are actually being used and what kinds of results are obtained from the application of these techniques.

There are many references that propose methodologies. (Arquilla and Davis, 1994; Bacsi and Zemankovics, 1995; Balci, 1994; Davis, 1992; Gass et al., 1991; Hodges and Dewar, 1992; Kleijnen, 1995; Law and Kelton, 1991; Landry and Oral, 1993; Naylor and Finger, 1967; Sargent, 1994; Schlesinger et al., 1974; Shannon, 1975; Susceptibility Model Assessment and Range Test Project (SMART), 1995; Zykov, 1987) In order to cover the full spectrum of methodologies, but still keep the analysis to a readable size, four references were used as the basis of this analysis. Works by Balci (1994), Law and Kelton (1991), Sargent (1994), and Davis (1992) are used as the primary basis of the analysis because they appear to cover the full breadth of validation methodologies that have been espoused to date.

2.2 Validation Methodologies

The following is a description of the validation methodologies presented by Balci, Law and Kelton, Sargent, and finally Davis.

2.2.1 Osman Balci (1994)

The title of Balci's work, *Validation, Verification, and Testing Techniques Throughout the Life Cycle of a Simulation Study* (1994), implies that validation is not just a procedure to accomplish after a model has been constructed, but rather validation is a process that should be implemented throughout the entire life of the simulation. This concept of validation throughout the entire lifecycle of a simulation is an idealistic principle that is discussed in more detail later.

Balci's methodology is based upon six principles. The principles are:

1) Validation is not a 'yes or no' question.
2) Model validation should be conducted throughout the lifecycle of the model.
3) Validation requires independent analysis to prevent any biases of the model developer from entering the analysis.
4) Validation requires creativity and insight into the problem facing the analyst.
5) Complete testing of a model is not possible.
6) Validation must be planned and documented.

Balci, like most authors of simulation methodology, considers validation, and simulation work in general, to be an iterative process. Figure 2-1 shows Balci's representation of the entire lifecycle of a simulation study. The iterative property of the validation cycle means that the process is not strictly a sequential set of steps to perform.

Balci states that it is expected that the analyst may have to revisit a previous step should an error be discovered. Note that in Figure 2-1, validation is included in each reference to VV&T (Verification, Validation and Testing).

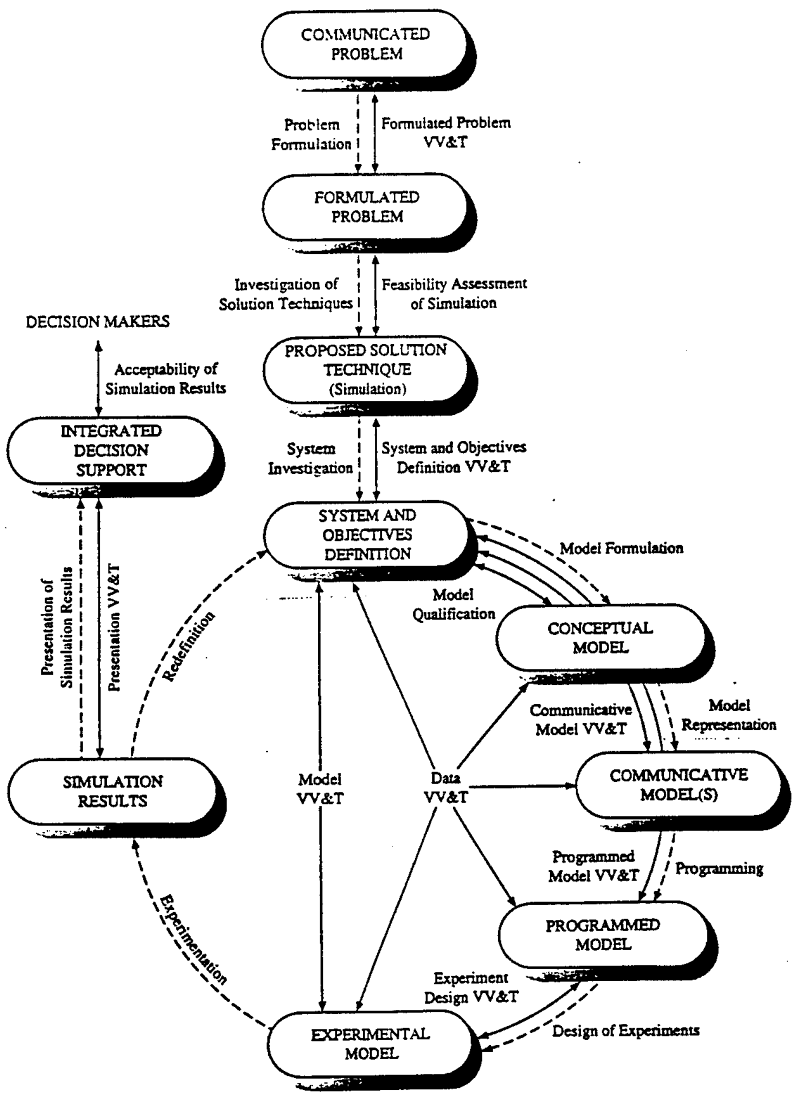

Figure 2-1: Balci Modeling Lifecycle

The areas of validation shown in Figure 2-1, referred to by VV&T, define Balci's methodology for validation. Explanation of those areas are as follows:

1) Formulated Problem VV&T. Formulated Problem VV&T is Balci's first phase of validation process. In this phase, the model has not yet been created. Formulated problem validation is the process of determining if the problem formulation is identical to the actual problem. If the formulated problem does not contain the actual problem, the analyst has committed a type III error, solving the wrong problem. (Balci, 1994) At this stage of the analysis, simulation has not been chosen to solve the problem. During the investigation of solution techniques, the analyst chooses the proper technique to solve the formulated problem. If simulation is chosen, the analyst continues along the lifecycle chart.

2) System and Objectives VV&T. System and Objectives Definition VV&T is the determination of the system characteristics for inclusion in complex system definition and modeling. This process is used to validate six major system characteristics that tend to cause many failures (Shannon, 1975):

1) System changes. The state of the system is an integrated result of past state changes and the basis for future states.

2) System Environment. All systems have their own environment and are part of a broader environment.

3) Counterintuitive behavior. Obvious solutions to discovered problems will often be ineffective in complex systems, because the cause and effect relationship of the problem might not be closely related in time or space.

4) Drift to low performance. Complex systems tend towards conditions of reduced performance over time.

5) Interdependencies. Complex systems have events that are influenced by their predecessor and affect their successors.

6) System organization. Complex systems usually exist in some type of organized state.

3) Model Qualification. Model qualification is the process of justifying the appropriateness of the conceptual model. Balci defines the conceptual model as the model formulated in the mind of the analyst. Model qualification is the process of justifying the assumptions that the analyst has postulated for the model.

4) Communicative Model VV&T. The communicative model is the model representation that can be communicated to other people. The communicative model can be judged against the real-world system, the study objectives, and the study constraints. Communicative model validation is the process of validating this version of the model as a proper form of the conceptual model.

5) Programmed Model VV&T. The programmed model is the computer executable code. This section could also be more appropriately called programmed model verification. This area does not include validation as defined in this work.

6) Experiment Design VV&T. Experimental design is the process of creating the experiments, or scenarios, with which the model will be exercised. Validating the experimental design is to evaluate the appropriateness of the scenarios for use to achieve the goals of the simulation analysis.

7) Data VV&T. Data validation is the process of checking that the input data is accurate, complete, unbiased, and used in the proper context for the model.

8) Model VV&T. Model validation is the process of checking that the experimental model is appropriately accurate to fulfill the study's objectives. The experimental model is the programmed conceptual model coupled with the designed experiments and the valid data.

9) Presentation VV&T. Presentation validation is the process of justifying that the output results are interpreted, documented and communicated with appropriate accuracy. Documentation is an extremely important factor in presentation validation.

As noted earlier in Balci's flow chart, simulation development, and specifically the validation effort, is an iterative process. The analyst continues to perform iterations of the steps of the validation process until the study objectives are met, or until the objectives are determined unattainable.

2.2.2 Law and Kelton (1991)

Law and Kelton suggest a three step approach to validation. As a preface to the statement of their methodology, Law and Kelton describe several general, but important, guidelines for validation that are not explicitly defined in their methodology.

1) Careful inspection and definition of the problem are required.

2) Expert analysis and sensitivity analysis should be used to determine the level of detail that the model requires.

3) Time and money constraints may be important and need to be investigated.

4) Real world systems that have a large number of factors require the use of a 'coarse' simulation or an analytic model to determine which factors are

important before a full-scale simulation model is developed.

5) Documenting model assumptions in a log is important to complete during creation of the model. Many assumptions may be forgotten by the end of the project.

The first step of the methodology is to develop a model with high face validity, meaning, the model looks reasonable to system experts. (Law and Kelton, 1991) The second step of the methodology is to test the assumptions of the model empirically. The third step of the methodology is to determine how representative is the output data from the simulation in relation to the real system.

In order to achieve high face validity, Law and Kelton suggest the following activities:

1) Conduct in-depth conversations with system experts. The process of collecting all of the information from the different experts can be valuable in its own right, regardless of the simulation study performed. It is rare that one person is extremely familiar with the entire system. Bringing together all the relevant information can be a beneficial excersize. (Law and Kelton, 1991)

2) Collect any data, historical or otherwise, that is from a system similar to the one in question. A system that is similar to the one in question can be used for data collection to help build the model. The analyst must be very careful to make certain that the data is representative and correct.

3) Use established, relevant theories. Well known, documented theories can be used to ease the modeling process. For example, the interarrival rate of customers to a

service system, such as a bank, is likely to be an independent, identically distributed (IID) exponential random variable.

4) Use relevant results from similar simulation models. Results from studies that contain some of the same characteristics or scenarios can be used.

5) Use experience and intuition. Using experience and intuition seems fairly obvious, but it is sometimes necessary to make assumptions for models that are based on experiences and intuition. Analysts can sometimes use knowledge gained from unrelated models.

6) Keep continuous interaction with the customer/client throughout the study. Interacting with the customer/client can clarify the problem. Interacting also keeps the client interested and involved in the process. This interaction can increase the validity of the model, since the client is generally the person who knows the most about the system. The client will understand the results better as well as be more confident in the study if he or she is involved throughout the developement of the model.

7) Perform a walk-through of the conceptual model to all key people. Before coding begins, a walk-through of the conceptual model will help validate the analyst's conceptual model and assumptions.

For the second step, Law and Kelton suggest empirically testing the model assumptions. Many techniques exist that can be used to test the model assumptions. (See Appendix A of this thesis for detail concerning techniques.) Law and Kelton suggest: 1) testing the probability distributions used, and 2) sensitivity analysis on output data.

The third step, and most important according to Law and Kelton, is the determination of how closely the model output data resembles the expected (real-world) output data. Many techniques are available that can be applied for testing output, depending on the situation. These techniques include, but are not limited to, Turing Test, Animation, and Time-series analysis and other statistical analysis techniques, etc. (See Appendix A of this thesis for detail concerning techniques.)

A specific point to which Law and Kelton call attention, that is not in the other references, is the use of a calibration factor. When model outputs do not agree exactly with real system output data, often times a calibration factor is either added or multiplied to possibly achieve the correct absolute output. Law and Kelton stress that caution must be taken when using a calibration factor. A calibration factor may achieve proper results for one set of input data, but the model might not be valid over the entire range of inputs. A possible solution to this problem, presented by Law and Kelton, is to use one set of data to create the calibration factor and an independent set of data to validate the use of the factor.

Law and Kelton's three step approach to validation is a mix of empirical tests, subjective tests, and common sense. Law and Kelton stress that empirical tests of output data are the most definitive tests for validation. This three step approach is based on the process outlined by Naylor and Finger (1967) in *Verification of Computer Simulation Models*. Naylor and Finger's work is recognized as one of the original important achievements in simulation validation. (Law and Kelton, 1991)

2.2.3 Sargent (1994)

Sargent proposes three approaches to validating a simulation model. The first approach, and most commonly used, is for the model development team to test the conceptual model themselves, and decide if the model is valid.

The second approach employs an independent validation team (or third party validation) to validate the model. This approach eliminates biases that may be inherent in the model developer, because someone removed from the original model development conducts the process of validation. The model developer would still be needed to guide the validation team through the model, but the developer is obliged to convince other simulation experts that his model is correct for the problem. A drawback of this effort is increased expenditure of time and money, since independent validation generally takes longer to complete than a similar effort by the model developer. A variation of independent validation is to have an independent team review the validation effort made by the developer. Review by the independent team would ensure a proper effort was made, but would not take the length of time required for the team to become completely knowledgeable in the model.

In order for an independent validation team to carry out the validation effort, extremely detailed documentation of the modeling effort must be available. Documentation of the entire modeling effort, especially the validation portion, is an important aspect of the validation process. Documentation should be a common sense procedure. (Shannon, 1975) However, under time constraints, documentation is one of the first things in an analysis study that is dropped. (Davis, 1992) Lack of proper

documentation is apparent when inspecting DoD combat models. Many such models are inadequately documented. (Davis, 1992) The Defense Modeling and Simulation Office (DMSO) has acknowledged the need for proper documentation and has created guidelines stressing the importance of good documentation. These guidelines are listed in Davis (1992).

The third approach uses a scoring model. The validating analyst assigns a score to each validation test performed to measure the effectiveness of that test. Scores are determined subjectively by the analyst when conducting the various techniques in the validation process. The scores are weighted and combined to form an overall score. The model is declared valid if the overall score surpasses a minimum passing score. A scoring model sounds like a good tool, but actually has several negative features. The subjectiveness of the scoring process can become hidden behind a seemingly objective score. The score can also cause overconfidence in the model. A model could possibly pass and still have a large deficiency in one or two areas. Lastly, who is to determine what score is passing or failing? The choice of the passing score adds more subjectivity to the analysis. The use of scoring models in validation is not used in practice very often. (Sargent, 1994)

Figure 2-2 is a visualization of Sargent's modeling process, including validation. As seen in Figure 2-2, Sargent presents his validation effort as an iterative process combining data validation, operational validation, and conceptual model validation. The validation elements of Sargent's modeling process are as follows:

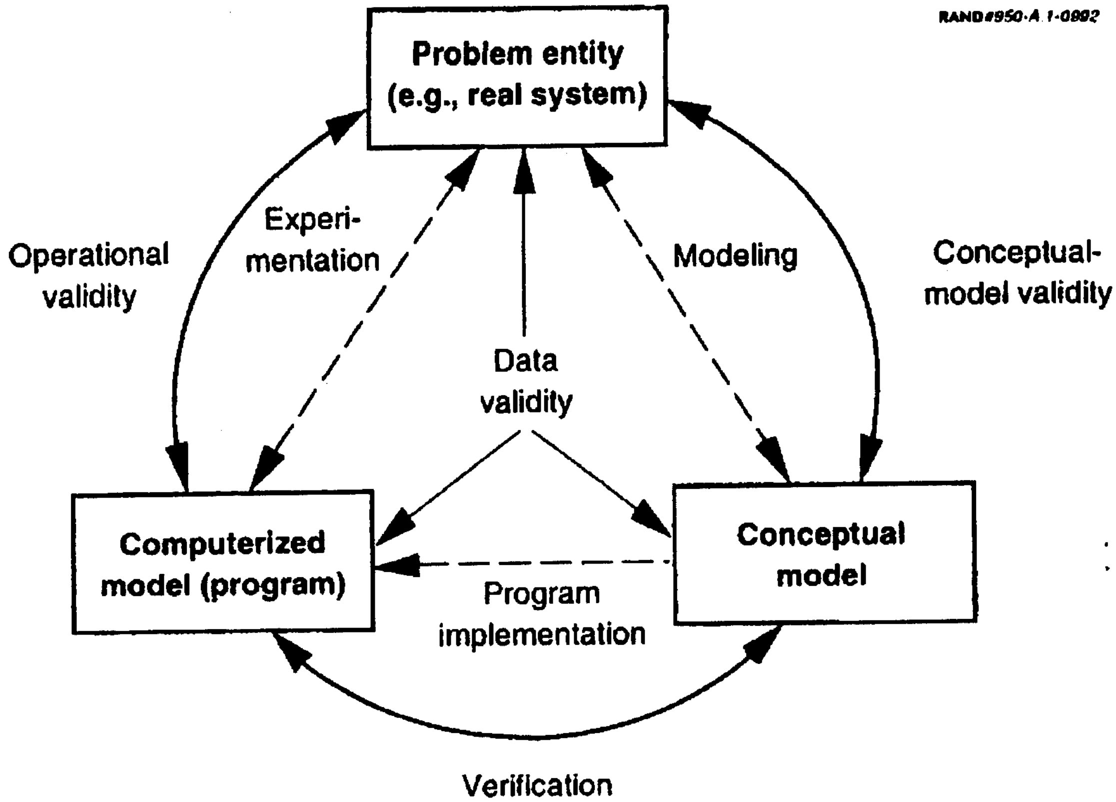

Figure 2-2: Sargent's Modeling Process

1) Data Validation. Although Sargent asserts the need for data validation, he declares that there is not a lot that can be done to ensure valid data. The best that an analyst can hope to do is develop good practices for collecting data and test the data for outliers and consistency. (Sargent, 1994)

2) Conceptual Model Validation. Conceptual model validation is the process of examining and justifying the theories and assumptions used in a model.

3) Operational validation. Operational validity is concerned with determining whether or not the model is appropriate for its intended purpose. This area is where the majority of Sargent's validation occurs. Three basic techniques are the most commonly

used for comparison between model and system data: 1. Graphs, 2. Confidence intervals, 3. Hypothesis testing.

As a culmination of his validation discussion, Sargent proposes a methodology to use as a minimum set of procedures for a validation effort.

1) The analyst and customer should agree before the study begins on the basic validation approach.

2) The assumptions and underlying theories of the model should be tested.

3) Face validation of the model should be checked on the conceptual model with each model iteration.

4) The model's behavior should be checked with the computerized model on each iteration.

5) The analyst should compare the model and system behavior for at least two sets of experimental conditions.

6) The analyst should fully document the validation process.

7) Schedule periodic reviews of the validation, if the model will be used over time.

2.2.4 Davis (1992)

Davis classifies the validation process into three general categories of validation techniques: Empirical Evaluation, Theoretical Evaluation, and Evaluation by Comparison. The three categories of techniques are used to achieve three types of validity: descriptive, structural, and predictive.

Descriptive validity refers to the model's ability to explain phenomena. Descriptive validity is an explanation of the model's capability to describe why an event occurs, and what events transpired beforehand to cause this event occurrence. Structural validity

means that the simulation has the appropriate objects, variables and processes modeled correctly for the simulation's needed use. Predictive validity means that the model can effectively predict the desired response of the system, at least within the domain of the specific initial conditions.

Davis suggests a potential list of techniques to use for validation. These techniques are described in detail in Appendix A. Davis' methodology is very heavily weighted on the use of face validation. He states that most serious errors in models are detectable through proper face validation. He does include a caveat by warning that the dangers of depending only on face validity are 'obvious', in other words, validation based solely on face validation is a bad idea. The dangers can be minimized if the validation effort includes a very broad face validation and in-depth spot checks using empirical tests. These checks are carried out via the empirical methods listed in Appendix A of this paper.

As a large portion of Davis's methodology, proper face validation requires several prerequisites:

1) The model is well documented

2) The model reviewers are familiar with a good set of standard scenarios.

3) The model has output that is sufficiently aggregated to permit comparison with a familiar set of metrics.

4) The model should have easy access for spot check requests.

The following list is a summary of Davis' methodology:

1) Apply the definitions and concepts to communicate the important issues of VV&A.

2) Use empirical and subjective tests throughout the entire model.

3) Consider the costs of fulfilling validation requirements.

4) Data validation.

5) Explain the process to the customer.

Successful completion of this methodology results in declaring a model descriptively, structurally, and predictively valid. Partial successful completion can result in declaring only one or two of the three types of validity.

2.3 Comparison of Methodologies

The following section will compare and contrast the four validation methodologies just presented. The evaluation is intended to compare the methodologies at a broad level of detail. The aim of this comparison is to firmly delineate the general differences between the methodologies.

Balci's paper is the most comprehensive and detailed material of the four works. Balci separates the validation effort into eight specific types of validation defined at different times of the lifecycle. Law and Kelton present a methodology for validation that is more general in nature than Balci's, however, Law and Kelton's methodology is similar to Balci's in that they both stress empirical testing as the primary means of achieving validation. The two works include subjective techniques, primarily face validation, as an important, but secondary tool. The general aspects of Sargent's methodology are similar

to Balci's and Law and Kelton's work in that Sargent also stresses empirical tests as the most important.

Sargent's seven step procedure is more specific than Law and Kelton's three step procedure, but it is less detailed than Balci's methodology. Sargent's recommended methodology has several specific tasks, and several common sense procedures to perform in order to achieve model validity. This general approach is similar to the intent of Law and Kelton. Balci does not recommend specific tasks, rather Balci recommends many techniques to use to achieve positive validation in the particular validation classifications.

Davis' methodology shows a stark contrast to the other methodologies by emphasizing the use of face validation as a broad validation check and empirical spot checks of important factors. Davis states that rigorous empirical testing of the entire conceptual model is usually not possible, because of time and resource constraints.

All four authors accentuate the necessity of performing the validation process throughout the entire lifecycle of the model. This concept is fine in an ideal setting, but there are many finished models, especially military combat models, that do not have any documented validation completed. (Davis, 1992) None of the four authors explicitly address the issue of validating a model that has already been completed. Davis implies that the methodology could be adapted to use on a completed model, but that is all. One could surmise the effort required to validate an existing model specifically, but there was no formal documentation concerning this area in any of the references.

Davis uses Sargent's lifecycle flow chart (Figure 2-2) as an example of an idealized, not realistic, modeling process. Davis claims that in practice this process breaks down for several reasons.

1) Most organizations do not have the discipline to have serious design before letting the programmers go to work writing code. Davis states that this attitude results in unintelligible models. Programming before validation of the conceptual model goes against every author's views presented in this thesis. If an organization lets programmers start coding a model or sub-model before the conceptual model (or conceptual sub-model) is formally created and validated, no confidence can be placed on that model. The idea of creating a solution before proper problem definition is completed is wide spread in American engineering society. (Wedberg, 1990) Unsuitable model synthesis is often caused by the fact that projects are too dedicated to a timeline instead of quality work. Managers get too worried in producing results that make the project look good at the time without concern for future problems. (Nicholas, 1990)

2) The ideal structure breaks down due to the increase in technology. Simulation programs are becoming more advanced and much easier to use so that analysts can create the conceptual model with the software's user interface. Davis implies that an analyst can create a programmed model without first creating a conceptual model. It should be obvious to experienced simulation analysts that this idea is not possible in a good analysis. The conceptual model may not be documented on paper, but it exists in the analyst's mind. The conceptual model is then created on the computer. This conceptual model must still be validated as any model on paper must be validated.

3) Analysts' conceptual models are often vague and programmers have to fill in details, thereby defining the model. Once again, this should be obvious to experienced analysts that this reason is just the result of a poor analysis. All four of the authors recommend reviewing the conceptual model with the owners/users of the real system before programming begins in order to subjectively validate that the model is a good representation of the real system. This type of review would help ensure that the conceptual model contains enough detail of the real system.

Table 2-1 is designed to be an easy reference to the different authors' methodologies. The table will be used as a guide for synthesizing the authors' works into one methodology. The synthesized methodology will be used in the following chapters for comparison to military policy and case study methodologies.

Table 2-1

Validation Methodology Quick Reference	
Balci	**Law and Kelton**
1. Formulated Problem Validation.	1. Model with high face validity.
2. System and Objectives Validation.	2. Test assumptions.
3. Model Qualification.	3. Test output empirically.
4. Communicative Model Validation.	
5. Experiment Design Validation.	
6. Data Validation.	
7. Model Validation.	
Sargent	**Davis**
1. Specify effort with customer.	1. Apply definitions & concepts of VV&A.
2. Test assumptions.	2. Use empirical & subjective tests.
3. Examine face validity each iteration.	3. Consider costs of validation requirements.
4. Explore model behavior.	4. Data validation.
5. Compare model & system output (2 sets).	5. Explain the process to the customer.
6. Document.	
7. Schedule periodic reviews.	

2.4 Validation Techniques

The methodology used in the validation effort is the important aspect of how to perform a validation effort. There are many techniques that can be used in each area of validation. This section presents a brief discussion of the techniques suggested in the different references.

Table 2-2 shows a compilation of techniques documented for use in the validation methodologies by Balci, Law and Kelton, Sargent, and Davis. The techniques are sorted under two categories, Subjective and Empirical. The techniques listed under the Subjective heading require the analyst's judgment to decide the end result. The tests do not involve a mathematical conclusion. The techniques listed under the Empirical heading require objective analysis. Each empirical test requires experimentation and has a metric with which a conclusion is defined. The techniques are combined here into two categories for easy application. The techniques are described in detail in Appendix A of this thesis. Readers who are interested in the actual application of those techniques are referred there.

Table 2-2

Validation Techniques	
Subjective (Informal)	Empirical (Formal)
Face Validation	Statistical Analysis
Expert Opinion	Lab data
Doctrine	Historical data
Other Sources	Field test data
Analytic Rigor	Sensitivity Analysis
Comparison to valid models	Stress Test
Clarity and Economy	Black-box test
Relevant verisimilitude	Time-series Analysis
Experience/Intuition	Correlated Inspection
Existing Theory	Graph Analysis
Similar systems	Cause/Effect Graphing
Animation	Path Analysis
Walk-Through	Constraint Test
Formal Review	Inductive Assertions
Inspection	Proof of Correctness
Turing Tests	Traces
Event Validity	Extreme Condition Tests
Historical Methods	Fixed Values
	Predictive Validation
Peer Review	Internal Validity
	Historical Data Validation

2.5 Confidence: Value vs. Cost

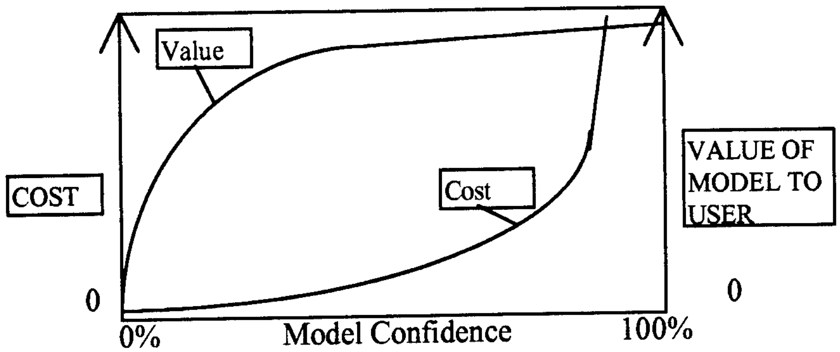

Figure 2-3: Cost and Value of Validation compared to Confidence

Certainly, one of the most important factors affecting an analysis study is cost. The graph in Figure 2-3 (Sargent, 1994) shows the relationships between the model's confidence, the value of the model's confidence, and cost associated with gaining the confidence level. Clearly, the exact design of the graph is dependent on the project situation. The amount of available resources, the time involved, and the nature of the project among others, will all have an effect on the shape of the two curves. Figure 2-3 is presented to show the general idea of the type of tradeoff associated with a validation effort.

All model validations will generally reach a point that it would require a large amount of resources (money) to increase the confidence a small amount. At the beginning of the analysis, a larger gain in confidence is achievable with a similar expenditure of resources, which results in a larger gain in value of the model to the user. The value of the model to the user will increase along with the model confidence, but the value gained will tend to level off as the confidence approaches 100%. This is better known as the Law of

Diminishing Returns. Cost usually becomes a significant factor in models that require high confidence, because of the potential consequences of invalid model results.

(Sargent, 1994)

The tradeoff between cost of confidence gained and value of confidence gained is an important aspect of validation because of the current trend of shrinking defense budgets. The resulting question that comes out of the cost versus value tradeoff is; Is the value of the model gained significant enough to warrant the expenditure to increase the confidence?

2.6 Methodology synthesis

Table 2-3 is this author's proposed methodology. The methodology is synthesized from concepts from Balci, Law and Kelton, Sargent and Davis, to produce a procedure that covers the entire range of ideas from each authors' works. The methodology proposed here will be referred to as the Proposed Integrated Methodology (**PIM**) model.

Table 2-3

PIM Model
1. Apply the definitions and concepts to communicate the important issues of VV&A to the customers.
2. Determine tradeoff of cost vs. value of the confidence gained.
3. Document all work in validation effort.
4. Examine validity of data.
5. Develop the model with high face validity throughout the entire building process, with system experts, experience and intuition, and Peer Reviews.
6. Experimental design validation.
7. Test or verify the assumptions made in the conceptual modeling.
8. Test the model's output with empirical techniques, especially if historical data exists.
9. Explain the process to the customer.

The PIM model contains all of the basic aspects of Balci, Law and Kelton, Sargent, and Davis. The following section is an explanation of the importance of each methodology step.

1) Apply the definitions and concepts to communicate the important issues of VV&A to the customers. Understanding the concepts involved with VV&A is important for the customer. The methodology used in any problem solution must be questioned when important decisions are going to be made. This fact is especially important when the analysis is carried out by a contractor to the government. The military and government contractors do no have the same agenda, and therefore careful examination must be made of the contractor's work.

2) Determine tradeoff of cost vs. value of the confidence gained. The tradeoff between cost of validation and the value of the confidence in a model is an important factor that the analyst and customer must decide together. Different studies have different driving factors in this tradeoff. For example, studies of command and control require very high confidence in the results and therefore, the cost would probably be secondary. For non-mission essential studies, the cost might be the driving factor.

3) Document all work in validation effort. Documentation is discussed as important in each reference. As noted previously, documentation is especially important to military analysts for continuity, because of the high turnover rate of military analysts.

4) Examine validity of data. Validation of the input data used in a model is important to be sure that the data is accurate, complete, unbiased, and used in the proper context for the model.

5) Develop the model with high face validity throughout the entire building process, with system experts, experience and intuition, and Peer Reviews. Many serious errors in models are detectable through proper face validation. (Davis, 1992)

6) Experimental design validation. Experimental design is the process of creating the experiments, or scenarios, with which to use the model. Validation of the experimental design is to justify the appropriateness of the scenarios for use in the model.

7) Test or verify the assumptions made in the conceptual modeling. Validating the assumptions made during the creation of the model is discussed by all authors. The simulation analyst creates assumptions for the model that have to be proven valid.

8) Test the model's output with empirical techniques, especially if historical data exists. Empirical analysis of output data is an important step in all references.

9) Explain the process to the customer. Explanation of the process to the customer is an important step. Study and analysis is a customer-oriented (support) effort. If the customer (user) of the analysis does not understand the concept of the work, the results or recommendations will probably not be used effectively.

For completeness of the PIM model synthesis, Table 2-4 shows that the PIM model incorporates all of the various author's methodologies. Table 2-4 also acts as a means to visually compare the various author's methodologies. Table 2-4 shows each of the various authors' methodologies with step names as row titles, and each step of the PIM model as the column title. The intersection of each row and column is marked with '**Y**' denoting the step in the PIM model that incorporates the row element. *# denotes a

comment concerning the relationship between a step in a given author's methodology and the PIM model methodology, and are described below the table.

Table 2-4: PIM model compared to reference methodologies

PIM model step>>	1	2	3	4	5	6	7	8	9
BALCI *************************	***	***	***	***	***	***	*3	***	***
1. Formulated Problem validation	*1		*2						
2. Sys. & Obj. validation							**Y**		
3. Model Qualification							**Y**		
4. Comm. Model validation					**Y**		**Y**	**Y**	
5. Experimental Design validation						**Y**			
6. Data validation				**Y**					
7. Model validation							**Y**	**Y**	
LAW AND KELTON *************	***	*4	*5	***	***	***	***	***	***
1. Model w/High face validity					**Y**				
2. Test Assumptions				*6			**Y**		
3. Test Output Empirically								**Y**	
SARGENT *********************	***	***	***	***	***	***	***	***	***
1. Plan validation effort w/customer	**Y**	*7							
2. Test assumptions				*8			**Y**		
3. Face validity					**Y**				
4. Explore model behavior					**Y**			**Y**	
5. Compare model & system output								**Y**	
6. Documentation			**Y**						
7. Schedule reviews	*9								
DAVIS *************************	***	***	***	***	***	***	***	***	***
1. Communicate validation issues to user	**Y**								
2. Empirical & subjective techniques used					**Y**			**Y**	
3. Analyze cost requirements		**Y**							
4. Data validity				**Y**					
5. Explain analysis to customer									**Y**
PIM model methodology steps >>	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

*1 Formulated Problem Validation is concerned with determining if the proper problem is being analyzed. For this comparison, it is assumed that the proper problem has been realized and that simulation has been decided upon as the best solution method.

*2 Although Balci does not include documentation as an explicit step in his described methodology, complete documentation is recognized in the work as an essential element to a successful validation effort.

*3 Balci's methodology reaches a very detailed level. Balci's second, third, fourth, and seventh steps are all similar in that they attempt to validate the conceptual model of a given analysis, four different times in the process. Step seven of the PIM model appears to encompass these particular steps.

*4 Law and Kelton do not explicitly include the cost versus confidence tradeoff analysis as a methodology step, however, they do include the implication of time and cost constraints in their general principles for validation.

*5 Documentation is included as being important in Law and Kelton's general guidelines, but not explicitly stated in their methodology.

*6 Law and Kelton consider the use of particular data as an assumption in the model. Therefore, using representative (valid) data is important to the modeling process, and is checked under the assumption step.

*7 Sargent includes the cost versus confidence tradeoff as a subsection of planning the validation effort.

*8 Data validity is specifically discussed in Sargent, but not explicitly defined in the list of procedures from his methodology. Sargent states that there are not many ways to ensure that the data is valid, except for using good procedures for collecting data, examining outliers, and using consistency checks.

*9 Schedule periodic reviews. All four authors discuss the iterative nature of validation. It is assumed here that the validation process is an on-going process and is checked as needed.

In this chapter, four references on validation are analyzed. The different author's validation methodologies are synopsized. The methodologies are compared at a general level and the validation techniques that the author's describe are presented. The PIM model methodology is synthesized to incorporate the aspects of all four reference methodologies.

In the following chapter, the synthesis of methodologies is compared to the current Department of Defense policies on validation of simulation models. Published case studies of validation efforts are then analyzed to determine what types of efforts are actually being performed by analysts and how the efforts relate to the PIM model.

3. MILITARY VALIDATION POLICIES

Modeling and Simulation (M&S) use is growing in the military community. Military policy covering M&S is in the process of being formed. Department of Defense directive 5000.59, *DoD Modeling and Simulation Management*, is the foremost policy covering military simulation models. The directive instructs each of the military components to establish verification, validation, and accreditation policies. In response to this directive, the Army has created Army Regulation 5-11, *Army Model and Simulation Management Program*. The Navy has a Naval Operational Instruction, OPNAVINST, *Verification, Validation, and Accreditation of Navy Models and Simulations*, which was still in draft as of February 1994. The Air Force has created an Air Force Instruction 16-1001, which is also still in draft form. This section will start with analysis of the Air Force policy.

3.1 Air Force Policy

Air Force Instruction 16-1001, still in draft form, defines validation as "the rigorous and structured process of determining the extent to which a model and simulation accurately represents the real-world phenomena from the perspective modeling and simulation use." The instruction presents two types of validation: structural and output validation. Structural validation includes examination of all algorithms, assumptions, and the model structure, in the context of the problem. Output validation includes

examination of the degree to which the simulation results accurately compare to the perceived real world system.

The definition of validation in AF Instruction 16-1001 implies that the simulation must be comparable to the real-world system, which is restrictive, because some analyses only need relative differences, not absolute differences. (Kleijnen, 1995) Also, since a model always contains abstractions, there is no model that is perfectly valid. A possible change in this definition might be that validation determines if the model is good enough for use. Being 'good enough' is dependent on the goals of the analysis for which the model is being used. (Kleijnen, 1995) As stated in Chapter 1 of this thesis, the definition of validation used in this thesis is the process of determining if a conceptual model is suitable for use to achieve the goals of the particular simulation.

Air Force Instruction 16-1001 requires that a documented validation effort is made on models that fit any of the following criteria:

1) Engagement, mission, or any campaign level models that will be briefed to senior ranking officials outside of the Air Force;

2) Models used significantly in a cost and operational effectiveness analysis;

3) Models used for force structure, resources, warfare requirements, and assessment analysis;

4) Models used in Acquisition projects involving over $115 million in research, test, design, and evaluation or $540 million in procurement;

5) Models with 'real time' control and movement of troops;

6) Models with aspects dealing with human safety;

7) Models made available to agencies outside the Air Force that AF/XOM has determined warrant the attention.

These criteria apply to many projects, but not all. For simulation projects that are not encompassed by this list, the Instruction does not mandate validation, but rather leaves it to the decision of the respective MAJCOM. The Instruction does not mandate how to validate models, but rather defines a management policy for the VV&A process.

The Air Force Instruction does not mandate a methodology for use in validation of simulation models. However, a methodology for validation is implied in the definition of validation; structural and output validation.

1. Examine structural validity.

    - Make internal examination of simulation assumptions and algorithms.

2. Examine output validity.

    - Use empirical tests to determine how well the model results compare with the real-world results.

Although this list is not comparable to the methodologies listed in the previous chapter, some guidance for validation can be derived by analysts.

The A.F. Instruction does suggest a list of techniques for possible validation use. However, the instruction does not give guidance on how to apply the techniques and under what circumstances the techniques should be used. The list of techniques from A.F. Instruction 16 is as follows:

1) Face validation.

2) Comparison with historical data. (Statistical analysis)

3) Comparison to similar models already accredited.

4) Comparison to developmental test data.

5) Comparison to operational test data.

6) Peer review. (Subject matter experts analyze model and determine if it is an accurate representation of their system.)

7) Independent third party validation.

8) Threat data audits on models and simulations that are part of ACAT 1D and ACAT 2 programs that rely on threat data.

These techniques are all discussed under the description of validation techniques in Appendix A, with the exception of number 8. ACAT 1D and ACAT 2 programs are specific Air Force programs that are unreferenced in the Instruction. The actual process of data audit is similar to Balci's tracing. Tracing consists of monitoring the low level path of threat data as it transfers through sub-models in the simulation.

Lack of concrete guidance in the methodology of validation is a shortfall of the instruction. As noted in the SMART report, *Comparative Analysis of Tri-service Accreditation Policies and Practices* (1995), "the major shortcoming of the Air Force process is the lack of guidance on the criteria that should be used to determine the amount of VV&A required."

3.2 Army Policy

The Army covers Model and Simulation (M&S) management under regulation 5-11, entitled *Army Model and Simulation Management Program*. Regulation 5-11 is a Headquarters, Department of the Army document, covering the management of simulation models.

Army Regulation 5-11 defines validation as "the process of determining the extent to which M&S accurately represent the real-world from the perspective of the intended

use of the M&S." The regulation then states that the ultimate purpose of validation is to validate the 'entire system', which consists of the M&S, data, and the operator-analyst who will execute the simulation.

The Army definition is very similar the Air Force definition. The same argument used for the Air Force definition is applicable for the Army definition. Both of the definitions of validation imply that the simulation must be comparable to the real-world system. Since a model always contains abstractions, no model is perfectly valid. As with the Air Force definition, a possible change in this definition might be that validation is the process of determining if a conceptual model is suitable for use to achieve the goals of the particular simulation.

Like the AF Instruction 16-1001, Army Regulation 5-11 suggests possible techniques for validation use. These techniques are:

1) Face validation;
2) Comparison with historical data;
3) Comparison with other simulation results;
4) Comparison with engineering test data;
5) Comparison with operational test data;
6) Peer Review (face validation by system experts);
7) Independent or third party validation.

This list of techniques is essentially identical to the list in the Air Force Instruction. These techniques form a small subset of the validation techniques reviewed in Appendix A. Like the A.F. Instruction, the regulation does not mandate a specific procedure to

follow. The regulation does mandate the use of a systematic plan for VV&A of all Army models, but no plan in particular is specified. The requirement for a VV&A plan, and the subsequent lack of any guidance is, like the A.F. Instruction, a large shortcoming of the regulation.

3.3 Navy Policy

The Navy is in the process of creating policy covering modeling and simulation. The draft operational instruction is titled *Verification, Validation, and Accreditation of Navy Models and Simulations*.

The Navy instruction defines four levels of VV&A information requirements to cover all Navy models. The level of effort is dependent on the tradeoff between the risk of using an inaccurate model and the cost of validating the model to a higher level. Level 1 VV&A requires documentation of model development, improvements, past applications, any validation effort performed, and defines the application domain for use. Level 2 requires examination of the model's assumptions, algorithms, architecture, and implementation in addition to level 1 requirements. Level 3 requires analysis of the model's application results in addition to requirements for level 2. Level 4 VV&A is used for models of real time movement of forces or those that deal with human safety. Level 4 effort includes all of the requirements for lower levels, performed at an 'extraordinary level.'

One interesting note seting the Navy instruction apart from the other service policies is the requirement for an independent team to either 1) assess the validation work

performed by the simulation analysts, or 2) perform the validation work themselves. Independent verification of the VV&A work increases the probability of having a good model, but adds to the cost and extends the length of time needed to develop the proper tools for the analysis.

The instruction defines the information requirements for each level of effort, but does not define an acceptable level of effort for each requirement. The information requirements define a validation methodology as follows (VV&A level in parenthesis):

1) Design documentation. (levels 2,3,4)
2) Determine level of V,V & A needed by cost vs. confidence required. (all levels)
3) Summary of assumptions, algorithms, architecture, and data. (levels 2,3,4)
4) Face validation. (levels 2,3,4)
5) Comparison to real world data. (level 3,4)
6) Data validation. (levels 3,4)
7) Users and analysts trained. (level 3,4)

The seven steps are validation elements of the overall VV&A process.

3.4 Tri-service Policy comparison

After analysis of the three services' policies, the Navy seems to have the most guidance for validation. The Army and Air Force policies do not contain enough guidance on how to conduct a validation effort. The Navy draft policy on the other hand, defines a methodology. The Navy methodology is summarized in Table 3-1.

Table 3-1

Navy Methodology
1. Design documentation. (levels 2,3,4)
2. Determine level of V,V & A needed by cost vs. confidence needed. (All Levels)
3. Summary of assumptions, algorithms, architecture, and data. (levels 2,3,4)
4. Face validation. (levels 2,3,4)
5. Comparison to real world data. (level 3,4)
6. Data validation. (levels 3,4)
7. Users and analysts trained. (level 3,4)

The methodology implied by the Navy policy is extremely close in detail to the PIM model (Table 2-3). Several elements of the methodology comparison are worth noting. The first element of note is the cost tradeoff with confidence required. The Navy instruction highlights the cost versus confidence tradeoff of the validation effort as a key component.

The second element of note is the requirement for an independent team to either 1) perform VV&A on the model in question, or 2) examine and verify the VV&A effort performed by the analyst who created the model. Requiring the independent check will greatly improve the probability of a valid model and analysis, but can become costly for some analysis projects. Smaller projects might not be worth the cost of performing the independent check. In some cases, the analyst's VV&A effort might be suitable. Sargent (1994) states that independent VV&A is definitely too costly for the benefit gained. Sargent suggests that the independent VV&A only examine and verify the VV&A work completed by the analyst.

Table 3-2 shows the comparison between the Navy policy and the PIM model.

Table 3-2: PIM model compared to the Navy policy model.

PIM model step>>	1	2	3	4	5	6	7	8	9
1. Design documentation.			Y						
2. Analyze cost vs. % confidence.		Y							
3. Model documentation.			*1						
4. Face validation					Y				
5. Comparison to real world data								Y	
6. Data validation				Y					
7. Users and analysts trained									*2

*1 Summary of assumptions, algorithms, architecture, and data.

*2 The intent of the users and analysts trained validation step is that the model is only valid if the users of the model are trained properly. For this study, it is assumed proper training is given.

3.5 Conclusions on Policy

The Army has the regulation 5-11 concerning VV&A, while the Air Force and Navy have draft policies under creation for VV&A. In their present forms, the Army and Air Force policies have shortfalls. The two policies define how to manage simulation studies, but give no guidance on how to carry out the actual VV&A effort, specifically the validation portion. The Navy policy presents a validation methodology to be used. Inspection shows that the Navy methodology is extremely close to the PIM model.

By requiring a third party, independent validation, the Navy has projected itself as the most concerned over proper analysis. This type of concern over validation of models is very important, since modeling and simulation is becoming a more popular tool for use. Proper validation may cost more money in the analysis, but it can save much more money, and even lives, by catching mistakes in the simulation rather than in the real-world.

4. CASE STUDIES

This chapter contains evaluations of published case studies of validation efforts. Each case study evaluation will proceed by the following approach:

1) a description of pertinent model background information,
2) a description of the validation methodology used in the case analysis,
3) a comparison of the analysis methodology to the PIM model, and
4) an assessment of the shortcomings, benefits, and overall effectiveness of the methodology used in each case.

The PIM model is reprinted below for easy comparison to the methodologies used in each case study.

Table 2-3

PIM Model
1. Apply the definitions and concepts to communicate the important issues of VV&A to the customers.
2. Determine tradeoff of cost vs. value of the confidence gained.
3. Document all work in validation effort.
4. Examine validity of data.
5. Develop the model with high face validity throughout the entire building process, with system experts, experience and intuition, and Peer Reviews.
6. Experimental design validation.
7. Test or verify the assumptions made in the conceptual modeling.
8. Test the model's output with empirical techniques, especially if historical data exists.
9. Explain the process to the customer.

Six case studies covering a broad range of model topics, from a classified military model, to an ecological model of a fish habitat, were evaluated. However, the case studies were not chosen because of the broad range of their subject matter, rather, they represent the entire set of published, detailed validation efforts that could be found through an extensive literature search. This confirms Kleijnen's (1995) assertion that, "case studies on validation are rare."

4.1 Case Study 1: RETACT Model

4.1.1 Model Background

The Real-Time Advanced Core and Thermohydraulic (RETACT) nuclear power plant simulation (Balci, 1987) is a mini-computer based, real-time simulation model used for analysis of nuclear power plant control and engineering. Nuclear power plant simulations are normally run on large mainframes and do not operate in real-time. The simulation focuses on modeling the reactor coolant system thermohydraulics and core kinetics. Model validation is of obvious importance since failure of the real-world system could result in thousands of fatalities.

Six test facilities were built that enabled analysts to conduct a large variety of extremely detailed experiments that could not have been conducted in an operational nuclear power facility. The test facilities thus gave the analysts access to data that would otherwise not exist, especially data from experiments that studied the effects of power plant accidents. The use of the test nuclear facilities provided a better understanding of

the complex thermohydraulic processes and added credibility to subsequent analysis of the simulation results.

4.1.2 Validation Methodology

The simulation was created with forethought of the validation effort, specifically so that the model output could easily be compared to test data. The validation approach consisted of several empirical comparisons of model data to test and real world data. Statistical analysis was the primary technique used in conducting these empirical comparisons. As a secondary validation effort, the senior plant control operators of the nuclear reactor performed face validation on the model. Finally, data collection and manipulation processes were validated by the analysts and plant operators.

The following sequence of methods was used in the RETACT model validation effort:

1) Model development included plan for validation.
2) Face validation performed by system experts.
3) Data validation.
4) Model output tested empirically against test and real world data.

4.1.3 Methodology Analysis

Table 4-1 shows which components of the PIM model were used in the RETACT validation effort. The table contains a 'Y' in the row/column intersection where the PIM model contains the validation step performed in the case study.

Table 4-1: RETACT methodology compared to PIM model

RETACT model step>>	1	2	3	4
1. Communicate validation issues to customer.	*1			
2. Cost vs. value of confidence gained.				
3. Documentation.	*2			
4. Examine data validity.			Y	
5. Face validation.		Y		
6. Experimental design validation.				
7. Test assumptions.				
8. Test model output empirically.				Y
9. Explain process to customer.				

*1 It is implied from the documentation that the important issues of the validation effort were discussed by the analysts and plant operators, so that the analysts could create the model for easy comparison to test data.

*2 The documentation step is included because of the fact that the case study was published.

The PIM model steps excluded from the RETACT methodology are the cost versus value of confidence gained tradeoff, test assumptions, experimental design validation, and validation process explanation.

4.1.3.1 Face validation

The senior operators gave their opinion to the analysts that the model was indeed a suitable representation of nuclear power plant control. Furthermore, the analysts plotted the simulation output data and the test facility data in a time-series output graph and subjectively approved, by inspection, that the two series had sufficiently similar time-series patterns.

4.1.3.2 Empirical output analysis

The analysts used statistical analysis, primarily time-series analysis, to show that the model output was a valid representation of the real-world data. The analysts used data

from multiple model runs in statistical analysis tests with data from two test facilities and the real power plant. Readers who are interested in the details of the actual techniques used are referred to Appendix B, Section 1 of this thesis.

4.1.3.3 Data validation

Data for the simulation analysis was collected from the six test facilities as well as from actual nuclear power plants that had undergone major transients. The test facilities were scaled down versions of actual nuclear power plants. The data from the actual power plants was not accurate enough for sole use in the simulation, but it allowed the analyst to correctly scale the data from the test facilities (such as power output, volume of coolant, etc.). The analysts and senior operators subjectively validated that the scaling of the test data was a legitimate assumption.

4.1.4 Shortcomings, Benefits and Overall Effectiveness of the Validation Methodology

The validation effort gives no mention to any evaluation of the tradeoff between cost and confidence. The subject of the simulation, however, is important enough that the tradeoff would be heavily weighted in the favor of confidence. The validation effort did not explicitly test, or examine, all of the assumptions made in the model development. The analysts did subjectively approve the assumption concerning the data collection and manipulation as discussed previously.

The validation methods appear to have been useful in increasing confidence in the validity of the model. The subjective face validation effort increased the confidence in the model by the system experts affirming that the model was an accurate representation of

the power plant and that the data was properly collected and manipulated. The empirical statistical tests added confidence by showing a close match existed between the simulation data and the test data.

The analysts concluded that the RETACT simulation gives predictions for nuclear power plant operation as accurately as the mainframe based simulations, and has the added feature of operating in real-time. The analysts believe that their validation effort was extensive enough to declare the simulation sufficiently validated for use.

4.2 Case Study 2: HUNTOP Model

4.2.1 Model Background

The Naval mine hunting model HUNTOP (Kleijnen, 1995) was created to simulate the hunting of mines by ships using SONAR. SONAR propagates sound waves into the water, then detects the reflection of the sound waves off of objects, such as mines. The objects are detected by a human operator observing an echo that appears on the SONAR screen.

The simulation models an area of the ocean with randomly placed mines and other objects, that can be mistaken for mines. Simulated ships modeled with SONAR search for mines by tracing out sections of the simulated ocean. The mines can only be detected if they are in the small range of the SONAR. The key factors in detection of the mines are the SONAR window of illumination, the ship position, and the human operator. The position of the ship is a fairly obvious factor, since a ship can only detect mines if it is actually above the mine. The SONAR window is the area that the SONAR is illuminating

at any given instant. Finally, the human operator will always have a probability of error. Other factors that are not as significant in detection of the mines are the size of the mine, the echo created by the mine's environment (i.e., noise), the angle that the SONAR reaches the mine, and other acoustic noise created by other ships, waves, fish, etc.

One assumption made in the model concerns the Sound Velocity Profile (SVP) in the water. The SVP maps sound velocity as a function of the depth of the water. SONAR accuracy is dependent on the velocity that sound travels through water. The model uses a simple piecewise-linear SVP that remains constant throughout each simulation run. In reality, there are many factors that can change the velocity of sound through water. The analysts decided that the variations in the velocity were not significant enough to warrant the extra effort of creating a more accurate SVP.

A second assumption is with human behavior. The behavior of the human operator is represented by statistical distribution functions, called operator curves. Several curves give the probability of detection, which is modeled as an increasing function of the amount of time that the echo is visible on the screen.

The bottom of the ocean is modeled as a geometric pattern that is fixed for the length of each simulation run. Changing the ocean bottom pattern for different scenarios can add the uncertainty of nature to the simulation. Hills, valleys, and other prominent features of the ocean floor can hide mines, create SONAR noise, etc.

The model uses a parameter that does not have any physical interpretation to calibrate the results of the model, to coincide more closely to the real world results that

were observed in tests. The use of such a calibration factor was addressed by Law and Kelton and is discussed in Chapter 2, Section 2.2.2 of this thesis.

The field test data was collected from test runs of a SONAR equipped ship hunting a 'mock' mine field. Each mine location was marked on the SONAR scope. For purposes of the test, a mine was classified as detected only when a SONAR echo appeared in the marked area on the SONAR scope.

The HUNTOP model is intended to investigate different tactics for mine searching. Use of the model can help improve mine searching efficiency. The main intent of the model is to achieve relative results from different searching tactics of a particular mine field. A secondary objective is to achieve absolute predictions of mine detection probabilities for each search tactic. Scenarios can be set up for either approach.

4.2.2 Validation Methodology

This case study provides an example of independent, or third party validation. The HUNTOP model was a completed model when the validation effort began. Since the validation analysts were not involved with the model development, the lifecycle approach of validation is limited. Regardless, the methodology used by Kleijnen consists of essentially four steps. The first step in the validation methodology, model description, shows that the validating analysts had thorough knowledge of the model and system. The second step was face validation by system experts. Empirical testing of sub-models and the overall model using different statistical analysis techniques made up the third step.

Test data validation was the fourth step. Details of the techniques used in the HUNTOP validation effort are presented in Appendix B, Section 2 of this thesis. The following list summarizes the methods of Kleijnen's validation work:

1) Model description.
2) Face validation from 1) system experts 2) existing theory.
   -Including validation of assumptions.
3) Empirical testing of model output.
4) Data Validity.

4.2.3 Methodology Analysis

Table 4-2 is a comparison of the HUNTOP methodology to the PIM model.

Table 4-2: HUNTOP methodology compared to PIM model

HUNTOP model step>>	1	2	3	4
1. Communicate validation issues to customer.				
2. Cost vs. value of confidence gained.				
3. Documentation.	*1			
4. Examine data validity.				Y
5. Face validation.		Y		
6. Experimental design validation.				
7. Test assumptions.		Y		
8. Test model output empirically.			Y	
9. Explain process to customer.				

*1 The documentation step is included because of the fact that the case study was published.

The HUNTOP methodology lacks the PIM model steps of cost versus confidence tradeoff, empirical design validation, and any communication with the user or customer. Interaction with the customer may have occurred but was not documented. Although no

cost versus confidence analysis was presented, the cost of validation is probably secondary to the desired confidence, since the benefits acquired from use of the model can help save sailor's lives.

Kleijnen's validation relies primarily on statistical analysis. Kleijnen used several empirical statistical analysis techniques to test the relationships between the simulation output data and the test data. Kleijnen did not test the assumptions explicitly (such as the SVP, human operator curves, etc.), but it is implied that system experts validated the assumptions subjectively by inspection (face validation).

HUNTOP is made up of 40 sub-models. Kleijnen started the validation by examining the validity of the sub-models, then examined the validity of the entire model overall.

4.2.3.1 Sub-model Validation

Kleijnen used Response Surface Methodology (RSM) with sensitivity analysis as a large component of his validation work on the sub-models (see Appendix B, Section 2 for details). Subjective face validation was used, but was secondary to the empirical statistical analysis work done. Davis' statement, discussed in Chapter 2, Section 2.2.4 of this thesis, that it is infeasible (too long and too expensive) to conduct rigorous statistical validation on the entire model, is corroborated by the fact that because of time constraints, Kleijnen was only able to perform validation on two sub-models and a portion of the model overall.

4.2.3.2 Model Level Validation

Validation of the overall model was attempted by comparing real versus simulated probabilities of detection. Kleijnen obtained mixed results from the attempted overall

validation. Kleijnen compared the probabilities of mine detection from the simulation and the field tests from three different scenarios, each scenario with a different mine field layout. The comparison was made using hypothesis testing, and using a comparison of confidence intervals. The results from the hypothesis testing could not be printed due to the classified nature of the information. The confidence intervals, however, were presented. Kleijnen took each probability of mine detection from each of the three scenarios run, and created confidence intervals (with unreported confidence level) for each probability. A comparison was made between the intervals, shown in Figure 4-1. It should be obvious that the probabilities in scenario 1 have little chance of being equal. The probabilities in scenario 2 are very close to each other. The probabilities in scenario 3 are closer than scenario 1, but not close enough to have confidence in a conclusion of a valid model.

4.2.3.3 Data Validity

Kleijnen addressed the validity of the field-test data. Kleijnen did not directly question the validity of the test data, but implied that the field test could have been set up better and suggested the following refined testing procedure: Instead of only declaring a detection when a return is spotted in a circle on the scope, detections away from the drawn circles are assigned a probability of actually being a mine. The closer that the return is to the circled mine position, the higher weight it receives. For example, a detection 20 meters away from a circled position could be assigned a 90% probability of being a mine.

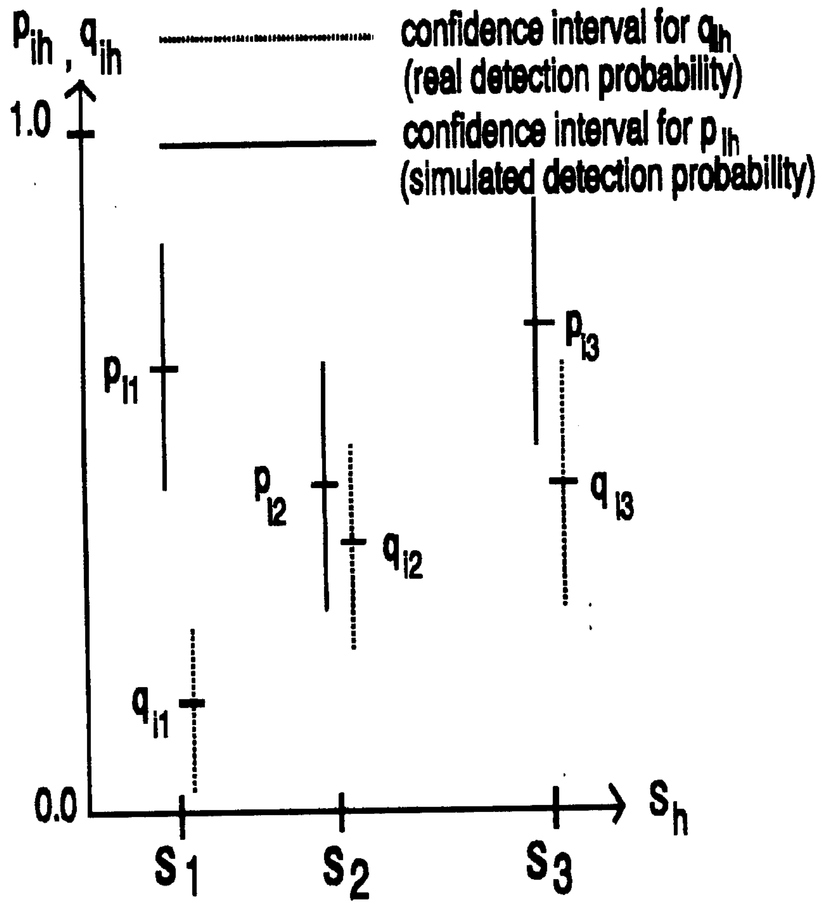

figure 1. Sensitivity of detection probabilities p and q to scenario S, for mine i

Figure 4-1: Comparison of simulated and real confidence intervals

4.2.4 Shortcomings, Benefits and Overall Effectiveness of the Validation Methodology

Time restrictions limited Kleijnen's validation effort to two of the forty sub-models. Confidence in the validity of the HUNTOP model could be increased by validating more of the sub-models in the model.

The documentation does not contain a description of interaction between the analysts and the customer. It is very possible that a large amount of interaction occurred, but the analysts might not have deemed the interactions worthwhile of case study documentation.

There was no reference to cost considerations in the validation effort. As stated before, this may have been due to the nature of the system being simulated.

Determination of the legitimacy of the calibration parameter is an important aspect of the validation that Kleijnen does not address. As stated in Chapter 2, Section 2.2.2 of this thesis, use of a calibration factor must be done carefully because the model might only be valid over a small range of inputs, and not the entire range of inputs. Kleijnen does not directly test the calibration factor as discussed earlier, but the three scenarios that were tested act to achieve the same result as directly testing the factor. Of the three scenarios that were tested by Kleijnen, only one set of the model and test data confidence intervals overlap significantly (Figure 4-1). This fact strongly suggests that the calibration factor made the model results look correct for the one set of inputs, but not for the entire range of inputs.

The methods that Kleijnen used appear to be beneficial to the validation of the model. Kleijnen's use of face validation by expert opinion for the RSM and sensitivity analysis for the sub-model validation appears to have incresed the confidence in the validity of the model. The analyst's determinations of which factors were important and which factors were not, agreed with the system experts' views. This determination led to a strong confidence in the sub-models that were tested. In the case of the detection

probabilities, the validation confidence was in the negative sense. A major drawback in the model (but a benefit gained by the validation methodology) was discovered by the confidence interval analysis of the probabilities of mine detection. This test showed that there is a validity problem with the model and that more testing is required.

Confidence in the validity of the model could be stronger if all of the sub-models had been tested. Kleijnen planned to validate all 40 sub-models using empirical statistical analysis, but was only able to validate two of the sub-models because of time restrictions. Kleijnen's validation attempt adds evidence in support of the statement by Davis that rigorous empirical analysis of a complex model is usually not feasible because it is too lengthy and expensive. Kleijnen's conclusion about the simulation based on the partial validation is that the model should not be used for prediction of future behavior, unless changes are made.

4.3 Case Study 3: RADGUNS Model

4.3.1 Model Background

The Radar-Directed Gun System Simulation (RADGUNS) is a simulation that models the detection, tracking, and firing performance of 20 different Anti-Aircraft Artillery (AAA) systems during engagements with several different types of airborne targets. RADGUNS simulates one aircraft versus one AAA battery. A secondary use of the model is evaluation of the performance of target aircraft with different characteristics against the AAA systems. Important system characteristics that the simulation encompasses include the weapon system, the operators, the target aircraft, flight paths, the

environment, and electronic countermeasures (ECM). Each weapon system is modeled with search and track radar, anti-aircraft guns, fire-control computer, servo aiming system, and the operational crew. The target aircraft's modeled characteristics include radar cross section, maneuvers, use of ECM, etc.

RADGUNS is a deterministic, rather than probabilistic, model. Making the model deterministic is an assumption that the real world system's characteristic relationships are well known and that their variability is very small, or that the variability of the system is not a concern for the study in question. Assessing the legitimacy of this assumption should be included in the validation effort.

4.3.2 Project Team

The Susceptibility Model Assessment and Range Test (SMART) project is located at the Naval Air Warfare Center at China Lake, California. The project team is part of the Joint Technical Coordinating Group. The SMART team was tasked by the Office of the Secretary of Defense to 1) develop a process for improving the credibility of simulations that are used in the acquisition of airborne weapon systems, 2) test the process on widely used models, and 3) expand the process to include all types of simulations. At this writing, all SMART project team documents are still in draft form.

4.3.3 Validation Methodology

The SMART project team created a methodology for verification, validation, and configuration management. The overall methodology was created from survey responses of modeling and simulation users and policy makers.

The validation methodology has three phases. Phase I is called Model Characterization. This phase consists of compiling background information to provide the model user with important information about the model. This information includes a synopsis with respect to the applicability of use of the model. The summary is directed towards answering questions the user might have about past model use, model documentation, model assumptions and limitations, and management of the model. This information is intended to be complete enough to let the user determine if the model is applicable for his or her intended analysis. Phase I could be looked at as an Executive summary for analysts. This information could help a user avoid a type III error, which is finding the solution to the wrong problem. (Balci, 1994) The primary purpose of Phase I for validation is to prepare the user for more rigorous validation in Phases II and III.

Phase II is a subjective review of the model structure and output by system matter experts. This effort is primarily face validation. This review by the system experts covers 1) validity of input data, 2) validity of the conceptual model, 3) all assumptions and limitations of the model, and most importantly, 4) sensitivity analysis of the model output.

Phase III of the validation effort is made up of detailed, empirical validation techniques used on the functional elements (sub-models) and the overall model. This process includes using statistical analysis to compare model results to real-world data gathered from operational or field testing, laboratory testing, and bench testing.

The following list is a summary of the methods employed by the SMART team:

1) A. Model Characterization
   - Model Use
   - VV&A history
   - Model management and support

   B. Model documentation
   - Assumptions and limitations
2) A. Subjective review by Subject Matter Experts
   - Face validation of model
   - Sensitivity analysis

   B. Data validation by subjective analysis
3) Detailed empirical testing (statistical analysis) of model results to test or real-world data.

4.3.4 Methodology Analysis

Table 4-3 shows a comparison between the SMART methodology and the PIM model methodology.

Table 4-3: SMART methodology compared to PIM model

SMART model step>>	1 A	B	2 A	B	3
1. Communicate validation issues to customer.	Y				
2. Cost vs. value of confidence gained.					
3. Documentation.		Y			
4. Examine data validity.				Y	
5. Face validation.			Y		
6. Experimental design validation.					
7. Test assumptions.			Y		
8. Test model output empirically.					Y
9. Explain process to customer.	Y				

The SMART methodology differs from the PIM model because it does not include a cost versus confidence tradeoff analysis. The lack of cost tradeoff is possibly due to the fact that the SMART project team was tasked to create their methodology for acquisition of new airborne weapon systems. Such a model would need extremely high confidence

because of both the very high cost of acquiring the real system, plus the fact that soldier's lives will depend on the new system. Model validation costs would be a secondary constraint behind confidence. The SMART methodology also does not include validation of the experimental design.

The validation of the RADGUNS model is separated into functional element validation and overall model validation. The methodology presented above was used on each of eight functional elements (sub-models) then on the model as a whole. The eight functional elements are flight path, target characteristics (radar cross section (RCS) static), waveform generator, thermal noise, angle track, range track, fire enable/disable, and ballistics. The validation effort was divided between the SMART project team and several different contractors. Specific details of the techniques used in the validation effort are listed in Appendix B, Section 3 of this paper.

4.3.4.1 Functional Element Validation

The model uses two methods to compute flight path information. The first method is computation by several subroutines in RADGUNS. The input data is manipulated by the subroutines and used by the model. The second method is computation by an external stand-alone program, called Blue Max, that manipulates the input data and creates an external file, which is then read by the RADGUNS model. The Blue Max data is used as a comparison tool for the RADGUNS data. Since the Blue Max program has already been validated, the data produced by the program is used to compare to the data created by the RADGUNS program.

The techniques used in the validation analysis of the flight path functional element of the RADGUNS model are statistical analysis, specifically the Mann-Whitney U test, and a subjective face validation of the standard deviations of the model and Blue Max test data sets. Visual inspection by the analysts of the data from the simulation and from the test data showed that the data sets are sufficiently close to each other. The use of the Mann-Whitney test showed fairly conclusively that the two data sets are from the same population, in other words the two methods of creating the flight path information are sufficiently identical..

The SMART analysts declared this functional element portion of the validation effort a success. All eight functional elements were validated in the same manner as the flight path functional element.

4.3.4.2 Model Level Validation

The model level validation was conducted with the same methodology as the sub-models. From inspection of the system and system expert advice, the SMART team determined that four applications of the model were the most important, and therefore were determined to be of primary interest in the overall validation effort. Those four areas are 1) target detection, 2) target tracking, 3) shooting performance, and 4) operator performance. Examination of these four areas of concern guided the analysts to the conclusion that validation of several of the concerns would be extremely difficult. Lack of credible test data, or unattainability of test data was the main reason for the analysts' doubt of acceptable validation.

Target tracking performance was one area that the analysts were able to examine in depth. Three sets of tracking error data from range tests were compared to two sets of simulated tracking errors. The results of the analysis were ambiguous. Some comparisons were favorable, and some were not. The comparisons were not conclusive in either positive or negative sense. The SMART analysts decided that more range tests needed to be conducted and that the comparisons would be continued with larger data sets. The analysts decided that the validity of the data was in question. Therefore, no conclusions pertaining to the validity of the model were drawn from the overall model validation effort. See SMART (1995, Accreditation Support Package) for full details of the techniques used, and the ambiguities discovered from the validation effort.

4.3.5 Shortcomings, Benefits and Overall Effectiveness of the Validation Methodology

The documentation does not include interactions with the customers. Interactions could have taken place, but they were not documented. The lack of cost and confidence tradeoff could be a drawback, except that the SMART project team was tasked to create a methodology and apply it to models that required very high confidence. Still, some mention of the cost tradeoff in the methodology would be beneficial.

As noted earlier, the documentation does not include justification of making the model deterministic, rather than probabilistic. Deterministic simulation models do not contain random variables, by definition. It seems that there would be many areas where random events could affect the performance of the AAA battery. Documentation of this decision would be appropriate.

The validation of the overall model led the analysts to discover potential problems with the model. The SMART analysts did not declare the model level validation effort a failure, but they concluded that the validity of the data was in question. The SMART analysts believed that their approach to the analysis was correct, but that the data used was not complete enough for proper comparison.

The model characterization phase produced detailed background information on the RADGUNS model. The documentation seems to be extensive enough to cover most conceivable questions about the model. Face validation (subjective) and statistical analysis (empirical) techniques used in the validation of the functional elements led to increased confidence in their validity. The model level validation proved to be effective by uncovering problems in the model.

The SMART analysis team declared the functional element validation effort a success and concluded that the elements were sufficiently validated. The analysis team did not make any conclusions concerning the overall model validity of the RADGUNS model, but decided that more testing was required.

4.4 **Case Study 4:** Star Field Model

4.4.1 Model Background

Star field simulations model different sections of the night sky to be used for the testing of navigation and tracking algorithms. Validation of the simulations historically has been very difficult because the sensors measuring real star fields produce massive amounts of data. Descriptions of the star fields include accurate spatial relationships as

well as the correct statistical distributions of number of visible stars. The simulation in Winter and Wisemiller (1974) models the space background and the sensors (Silicon Intensifier Target tube) used in the field's measurement. The output of the simulation is a 'mock' photograph of a particular star field to be used to test navigational equipment.

The key elements of the model for validation are 1) the reproduction of the position of catalogued stars, 2) the accurate modeling of the sensor image blooming, 3) modeling of the noise interference from background light, and 4) the creation of non-catalogued stars. The Smithsonian Astrophysical Observatory (SAO) catalogues stars that have an intensity magnitude greater than 9.5 (on a relative measuring scale, with no reported details). Image blooming is caused by saturation of a sensor element which causes spreading of energy into neighboring elements.

The main difficulty in validating a star field simulator is using data from real star sensors. The quantity of data from real sensors is of unmanageable proportions. One observation can produce one-third of a million light intensity values.

There is no catalogue of stars that are smaller in magnitude than 9.5. The simulation was created using an assumption that the exact locations of these smaller stars are not very important to the effectiveness of the simulation. A sub-model based on position in galactic latitude, randomly generates stars with magnitudes under 9.5 by position and magnitude.

4.4.2 Validation Methodology

Since the number of catalogued stars in any particular star field is known, the number of stars created by the simulation is easy to validate. Simply counting the stars of

magnitude 9.5 or greater, and comparing the total to the known number can partially validate the model. Star position and blooming are not quite so easy to validate. The analysts use statistical analysis to compare the relative geometry between three catalogued stars in each simulated field with the known geometry of the cataloged positions. Using relative positions as opposed to absolute positions of the stars is not as accurate overall, but the information flow is much more manageable. Image blooming is validated by comparing the statistical distributions of the sizes of the model output blooms to real sensor blooms.

The following list is a summary of the methods used in the star field simulation:

1) Statistical Analysis

4.4.3 Methodology Analysis

Table 4-4 is a comparison of the Star field methodology and the PIM model methodology. The documentation of the validation effort states that the effort relied solely on statistical analysis of the output data. The methodology used on the Star field model is limited for analysis.

Table 4-4: Star field methodology compared to PIM model

Star field model step>>	1
1. Communicate validation issues to customer.	
2. Cost vs. value of confidence gained.	
3. Documentation.	*1
4. Examine data validity.	
5. Face validation.	
6. Experimental design validation.	
7. Test assumptions.	
8. Test model output empirically.	Y
9. Explain process to customer.	

*1 The documentation step is included because of the fact that the case study was published.

4.4.4 Shortcomings, Benefits and Overall Effectiveness of the Validation Methodology

The documentation included description of empirical statistical analysis used in the validation effort. However, the documentation did not include details of the confidence levels used in the analysis techniques. There were no details documented concerning any subjective evaluation, assumption testing, communication with users, data validation, or experimental design validation.

Details of a cost tradeoff with confidence of the validation effort were not included in the documentation. The analysts do mention that validating the absolute position of the catalogued stars would be extremely difficult and time consuming. It is implied from this statement that the effort to perform absolute position validation would be too costly in time and resources. Details of this tradeoff would be beneficial.

In the documentation conclusions, the analysts seemed confident in the methodology used for validation. The analysts were satisfied that the results from the

statistical analysis techniques used to compare relative positions in the star field and analyze blooming effects were sufficient to declare the model valid for use.

It was noted in the documentation conclusions that there exists potential for inaccuracies of the measurements because of blooming. The potential was not great enough to warrant the analysts to perform more analysis on the data. Exact details of the inaccuracies would be beneficial in support of the validation methodology used.

4.5. Case Study 5: CERES-Wheat Model

4.5.1 Model Background

Zemankovics and Bacsi (1995) present a study of the validation of the simulation model CERES-Wheat. The CERES-Wheat simulation is used for crop growth analysis. The model incorporates several important factors in environmental management such as weather, soil type and characteristics, and management decisions. The model uses a data set that was the result of an initiative of the Technical University of Braunschweig. The data base was created to use as a common testing basis for various ecological models. (McVoy, et al., 1995) The database includes the following elements; soil type, daily weather data, nitrogen and water balance, and crop growth observations for three crops at three locations.

The model uses environmental information, farming management information, weather data, and six parameters that describe the type of wheat being analyzed as the model inputs. The model has many output variables but the main use is to predict how different types of wheat will grow under different conditions.

4.5.2 Validation Methodology

Statistical analysis of the output variables is the sole step of validation of the model. Five output variables were chosen in validating the CERES-Wheat model: the above ground mass of plant, leaf area, grain yield, and dates of antithesis and maturity. The principal techniques used were sensitivity analysis, confidence intervals and the t-test.

Sensitivity analysis was performed on two (P1D and G3) of the six wheat parameters, resulting in one of the parameters being much more sensitive to change than the other. Independent researchers claim that the three parameters, P1V, P1D, and P5, are more sensitive than the other three, G2, G3, and P5. Bacsi and Zemankovics found that P1D was much more sensitive than G3. Although this agrees with the independent research claim, Bacsi and Zemankovics did not make any claim about the validity of the model from that result. This comparison is implicitly face validation by system experts, but the authors did not comment on this result.

Full validation was not carried out because of the limited number of observations of the particular variety of wheat in the field data. The case study is presented as an excersize in the methodology used, as opposed to a full assessment of the model's validity.

The following list is a summary of the methods used in the validation effort:

1) Statistical Analysis of output data
   - Confidence intervals and t-test
   - Sensitivity analysis

4.5.3 Methodology Analysis

Table 4-5 shows the comparison between the CERES-Wheat methodology and the PIM model methodology.

Table 4-5: CERES-Wheat methodology compared to PIM model

CERES-Wheat model step>>	1
1. Communicate validation issues to customer.	
2. Cost vs. value of confidence gained.	
3. Documentation.	*1
4. Examine data validity.	
5. Face validation.	
6. Experimental design validation.	
7. Test assumptions.	
8. Test model output empirically.	Y
9. Explain process to customer.	

*1 The documentation step is included because of the fact that the case study was published.

The documentation included description of empirical statistical analysis used in the validation effort. There were no details documented concerning any subjective evaluation, assumption testing, communication with users, data validation, cost versus percentage confidence, or experimental design validation.

4.5.4 Data Validity

The test data set used in the CERES-Wheat model was not validated by Zemankovics and Bacsi, but the database has undergone extensive analysis for validity by other sources. McVoy, et al., (1995) demonstrate partial validation of the data using statistical analysis. McVoy references other works that also validate portions of the data set.

4.5.5 Shortcomings, Benefits and Overall Effectiveness of Validation Methodology

The methodology used in the validation was very limited. As noted previously, there were no details documented concerning any methodology steps other than empirical statistical analysis.

Zemankovics and Bacsi do not make any final claims about the validity of the CERES-Wheat model. The analysts state that the lack of field test data prevented them from performing a full validation effort. The data set used was not complete enough to gain conclusive results. In this type of simulation situation, Davis' methodology would be most applicable. With a general lack of suitable data for empirical testing, subjective validation techniques, as Davis suggests, would give the analysts the only validation results. However, subjective validation needs 'system experts' to give judgment for validation. If there are no such experts, the only validation possible would be from the experience and intuition of the analyst. It does appear that the authors could have made a subjective claim concerning the validity of the parameters as noted previously.

The authors concluded that larger data sets are needed for proper validation, which they did not have for the CERES-Wheat model. Zemankovics and Bacsi do not make any conclusions about the validity of the CERES-Wheat model.

Zemankovics and Bacsi did claim that the different statistical analysis tests would make for proper validation, given an acceptable data set. The authors state that assessing the validity of a model is often a subjective decision, and empirical tests can be useful to support such a decision. However, the authors do not make any subjective assessment themselves.

4.6 **Case Study 6:** Fish Habitat Model

4.6.1 Model Background

The Fish Habitat model was developed to predict the presence of cold water, cool water, and warm water fish in lakes of northern and southern Minnesota. (Stefan, et al., 1995) The simulation uses 25 years of daily weather data to model temperature and dissolved oxygen characteristics to be used as the main factors for the suitability to sustain fish life in 3002 lakes.

The simulation uses three variables to model the physical differences of the lakes: lake surface area, maximum depth, and the Secchi depth (the depth that a certain percentage of the radiation from the sun travels into the water). Several combinations of values of the three physical variables made for conditions that were unsuitable to support some of the fish. These cases were excluded from the analysis.

4.6.2 Validation Methodology

The methodology used in this case study is strictly statistical analysis. The sole test of reliability was comparing the simulated prediction of the presence of the different types of fish, to the observations of the actual fish populations at the respective lakes.

The lakes were first categorized by northern and southern portion of the state. The lakes were then separated into 27 classifications by all combinations of shallow, medium, and deep depth; small, medium, and large surface area; and eutrophic, mesotrophic, and oligotrophic Secchi depths. The number of northern lakes in each classification ranged from zero for shallow depth, large surface area, and oligotrophic

Secchi depth, to 531 lakes in medium depth, medium surface area, and mesotrophic Secchi depth. The number of southern lakes in each classification ranged from zero in shallow depth, large surface area, and oligotrophic Secchi depth to 168 in shallow depth, medium surface area, and eutrophic Secchi depth.

The model predicted the presence or absence of each type of fish for each classification of lake. If the prediction agreed with the observations, the class was labeled (A) for agreement. If the prediction and the observation did not match, then the class was labeled (D) for disagreement. The scores were quantified by assigning a 100% score to an (A) and a 0% to a (D). The scores were then averaged over the range of fish types for each classification to get a percentage for each (see Table 4-6 below).

The second comparison was performed by counting the number of lakes in which the most common fish was observed for each classification and reporting that number as a ratio of the total number of lakes in that classification. This ratio of lakes was compared to the simulated results of the percentage of lakes in each classification that are habitable.

The analysts examined the percentages of correct predictions and made a subjective assessment that the model did a good job predicting the presence of fish. There was no documentation of any other subjective assessments of the model.

Table 4-6 is the percentages of agreements between the simulation and actual observations.

Table 4-6

Simulation and Actual % Agreement						
	North			South		
	Cold	Cool	Warm	Cold	Cool	Warm
Fish presence	100%	100%	100%	18%	100%	100%
Most common species	16%	85%	59%	91%	85%	85%

The authors explained the disagreements observed in Table 4-6 by the fact that the suitability for fish existence is an average value and several fish, such as the cold water fish Cisco, are more tolerant to the temperature. The Cisco can survive in water at a deadly temperature for a short amount of time. Other explanations for disagreements include; human interference by stocking or eliminating fish species, winter conditions that were not modeled, uncertainty in measurements, among others.

The following list is a summary of the methods used in the validation effort:

1) Statistical Analysis

4.6.3 Methodology Analysis

Table 4-7 is a comparison of the fish habitat model methodology to the PIM model methodology.

Table 4-7: Fish habitat methodology compared to PIM model

Fish habitat model step>>	1
1. Communicate validation issues to customer.	
2. Cost vs. value of confidence gained.	
3. Documentation.	*1
4. Examine data validity.	
5. Face validation.	
6. Experimental design validation.	
7. Test assumptions.	
8. Test model output empirically.	Y
9. Explain process to customer.	

*1 The documentation step is included because of the fact that the case study was published.

The methodology used in the validation of the fish habitat model is strictly statistical analysis. As in the previous two case studies, there were no details documented concerning any subjective evaluation, assumption testing, communication with users, data validation, cost versus percentage confidence, or experimental design validation.

4.6.4 Shortcomings, Benefits and Overall Effectiveness of the Validation Methodology

The first point of discussion in the potential deficiency of the validation is the data validity. There are several areas of the data that the authors raise as potential cause for concern:

1) The 25 year data is averages of the seasonal maximum temperatures. There could have been one or two years of extremely cold conditions that could have killed off a fish species, but the average would not have been affected significantly.

2) The temperature and dissolved oxygen data have uncertainties in the measurements.

3) Some fish can survive short periods at a lethal temperature. The time of exposure to the lethal temperature may not have been long enough to kill off the species.

The next point of potential deficiency in the case study is the method of comparing simulated results to observed values. Simply comparing the percentages of correct predictions is rather simplistic. Furthermore, the subjective choice of what percentage is passing and what percentage would be declared invalid seems arbitrary and has no explanation in the documentation. The average of the 'fish presence' percentages in Table 4-6 is 86%, while the average of the 'most common species' percentage is only 70%. The analysts explain away these obvious deviations without commenting on the potential for an invalid model.

Fish should not be expected to be observed in all lakes that are suitable for their existence. However, the result of 18% for cold water fish in the southern portion of the state should show the possibility of an invalid conceptual model. The authors admit that more investigation is warranted to determine if more parameters than just temperature and dissolved oxygen should be used in the model, specifically for the southern cold water fish.

The analysts present the percentages of observations of the 'most common species' of each type (cold, cool, and warm) of fish, but there does not appear to be any benefit to the overall validation by this analysis. A fish species could be the most prevalent species in its particular type and still not be present in fifty percent of the lakes. The authors do not give statistics for the observations of presence of the most common

species. There does not seem to be any real comparison. The analysts seem to have committed a type III error, they answered the wrong question. (Balci, 1994) This validation process does not seem to answer the question of validity of the simulation model.

The analysts claim that the results in Table 4-6 show that water temperature and dissolved oxygen content in water are good indicators of suitable fish habitats. The simulation tries to predict which lakes are hospitable to the different types of fish. The output results show good agreement with the observed results in all of the 'fish presence' classes except for the southern cold water fish. It is apparent that more work is needed in the validation effort. The 'most common species' comparison does not seem to add any benefit to the validation. More testing is appropriate before the model can be declared valid enough for unlimited use.

4.7 Case Study Summary

Table 4-8 is a summary of the methodology steps from each of the six case studies.

Table 4-8: Summary of Case Study methodologies

Case study # >>	1	2	3	4	5	6	Total
1. Communicate validation issues to customer.	Y		Y				2
2. Cost vs. value of confidence gained.							0
3. Documentation.			Y				6 (*1)
4. Examine data validity.	Y	Y	Y				3
5. Face validation.	Y	Y	Y				3
6. Experimental design validation.							0
7. Test assumptions.		Y	Y				2
8. Test model output empirically.	Y	Y	Y	Y	Y	Y	6
9. Explain process to customer.			Y				1

*1 All six case studies include documentation, simply because of the fact that the efforts were published, except for case study 3, the SMART methodology, which explicitly recommends documentation of the modeling and validation process.

It appears from Table 4-8 that there is a minor disconnect between the validation methodologies that are created for publication and the actual validation efforts that are being performed by simulation practitioner. Assuming that the six case studies examined in this thesis are representative of actual practice, some conclusions about the types of efforts that practitioners find important and useful can be proposed. It is apparent from Table 4-8 that empirical analysis of output data is the strongest and most widely used method in validation. Face validation and data validation are secondary methods for use and the testing of assumptions and interaction with the customer are on the outer edge of apparent usefulness.

5. CONCLUSIONS AND RECOMMENDATIONS

This research examined the challenges that face military analyst in validating simulation models. The main challenges addressed in this thesis include examination and comparison of the following: 1) the types of validation efforts that academic simulation experts recommend as complete efforts, 2) military policy that guides the simulation analyst, and 3) actual efforts that simulation practitioners have performed.

This chapter starts with a brief review of the ideas on which Chapters 2, 3, and 4 are based. The remainder of the chapter is a discussion of conclusions from this research.

5.1 Summary

An extensive background research revealed a large quantity of references on the validation of simulation models. Four references were picked to represent the academic perspective on validation. The particular references were chosen because they appeared to span a large portion of the prevalent ideas on validation and the authors are well known and respected in the simulation field. This examination concluded with the creation of the Proposed Integrated Methodology (PIM) model, which is a synthesis of the ideas presented by the four references. The PIM model is shown in Table 2-3 in Chapter 2.

The military policy examination consists of analysis of draft Air Force instruction 16-1001, the draft Naval Operational Instruction, OPNAVINST *Verification, Validation, and Accreditation of Navy Models and Simulations*, and the Army Regulation 5-11, which

is the only military policy not in draft form at this writing. The policies are compared and contrasted between themselves and to the PIM model.

The analysis of published case studies was intended to determine the types of methodologies that analysts are actually performing. Unfortunately, only six detailed case studies were found after extensive research. This sample of case studies is too small to justify making broad conclusions, but they helped in gaining an idea as to the types of efforts that practitioners are actually completing.

5.2 Conclusions

5.2.1 Methodology Examination and Synthesis

There exists a very broad range of references concerning the validation of simulation models. Of the references examined in this research, many are very detailed in their approaches to validation, while others provide very general approaches. There does not seem to be one methodology that is accepted as best.

5.2.2 Military Policy

The Army and Air Force policies are very similar and share the same shortcoming: a lack of concrete guidance concerning the proper methodology needed to perform a satisfactory validation effort. Both policies describe the management of simulation models, but neither give specific direction with regard to the type of validation methodology that needs to be performed. More guidance in this area would be expected to produce better simulation studies.

The Navy policy gives much more direction than either the Army or Air Force policies. The Navy policy proposes a validation methodology to be used, that is very close in nature to the PIM model methodology. Although the Navy policy is still in draft form at this writing, it shows the most clear-cut direction for analysts to follow.

5.2.3 Case Studies

It is readily apparent from Table 4-8 of Chapter 4, the case study methodology summary, that there is a minor disconnect between the methodologies that the academics have created for publication and the methodologies that simulation practitioners are performing.

Assuming that the six case studies examined in this thesis are representative of actual practice, some conclusions about the types of efforts that practitioners find to be practical, important and useful can be proposed. It is apparent from Table 4-8 that empirical analysis of output data is the strongest and most widely used method in validation. Face validation and data validation are secondary methods for use and the testing of assumptions and interaction with the customer are on the outer edge of apparent usefulness and practicality.

There will exist some cases of simulation models where one method or another, such as empirical output testing, is the only means of achieving confidence in the model validity. However, the case studies examined show that increased confidence in the validity of the models was gained by the use of extensive methodologies, such as those including subjective assessments and examination of the validity of the data used.

For example, the CERES-Wheat model validation showed that strict use of only empirical techniques is sometimes not feasible. The analysts did not have appropriate data with which to validate the model empirically, and thus, were not able validate the model to any significant degree. A more extensive methodology might have helped gain some confidence in the model.

None of the case studies include explicit discussion of the cost of the validation effort versus the confidence gained from that effort. However, in the days of shrinking defense budgets, it is important for analysts to keep this trade-off in mind.

Time was a constraining factor in the HUNTOP validation effort, and can be viewed as a cost in the trade-off with confidence. The cost versus confidence gained trade-off depends on the simulation subject matter, the customer, and the analysts; but the trade-off is an important aspect of the overall model validation effort that needs to be addressed at the beginning of model development.

It is this author's conclusion that enough discussion has been presented to say that the PIM model is a reasonable methodology for use. The PIM model appears to be sufficiently realistic in scope so that it is feasible to implement. At the same time, it also appears to be extensive enough to guarantee an adequate validation effort. Although several steps of the PIM model were not shown to be used by practitioners, the ideas of interaction with the customer, cost versus confidence trade-off analysis and experimental design validation are important concepts to include in the validation process.

5.2.4 Observations

One observation that comes out of this research is that simulation practitioners and academic experts need to come together to find a methodology that can be acceptable to both. The academic experts need to research a methodology that incorporates the concerns of the practitioner, such as limited time and resources. Conversely, the simulation practitioners need to make a more concerted effort to conduct more extensive validation efforts on their models. Somewhere in the middle is a common ground where, hopefully, both can exist and simulation models can be economically validated.

A second observation is that because simulation analysis is intended to help decision makers make important decisions, it is vitally important that the simulation analyst performs a proper validation effort. If such an effort is not made, very little confidence should be placed in the analysis results and the effort expended to create the simulation would be for naught.

5.3 Recommendations for Validation Policy for Air Force Analysts

The following recommendations are presented to shape a policy that can help guide simulation practitioners in their work.

1) Use the unrestrictive definition of validation in the Air Force wide validation policy. Validation is the process of determining if a conceptual model is suitable for use to achieve the goals of the particular simulation.

2) Create a methodology for validation similar to the Proposed Integrated Methodology and define specific times throughout the lifecycle of the model development,

in the manner of Balci's lifecycle diagram (Figure 2-1), where milestones in the validation effort need to be achieved. Also, create a version of the methodology that can be used with completed models.

3) Define a set of levels, as the Navy Operational Instruction does, to categorize all models under Air Force control by level of importance and define an amount of effort that needs to be performed for each level of importance.

5.4 Recommendations for Further Research

The following recommendations are presented as potential topics for follow-on research to this thesis.

1) Examine the trade-off between cost of effort, the value of the model, and the percentage of confidence gained in validity of model in more detail. Determining how to approximate the cost of a validation effort and the value of the model after the validation is one potential area of research.

2) Research of more case studies of validation efforts would be beneficial. A larger sample size of validation case studies would produce a more lucid picture of what types of efforts that simulation practitioners are performing. The SMART project team will have completed documentation on the validation efforts that they performed on two more models, ESAMS and ALARM, as well as the completed RADGUNS effort, available for distribution as of 31 December 1995.

3) Research on the validation of distributed simulations is a topic that could be important in the near future. Use of distributed simulations is becoming more common.

Models that have been validated for working alone will now need some type of validation effort to achieve confidence to work in conjunction with each other.

Appendix A. Validation Techniques

Appendix A is a compilation of the techniques described by the various academic references to be used in each step of the methodologies.

Table A-1 is a summary of all of the validation techniques presented by the four authors. The techniques are categorized by subjective or empirical technique.

Table A-1: Validation Techniques

Subjective (Informal) Page 85	Empirical (Formal) Page 89
Face Validation	Statistical Analysis
Expert Opinion	Lab data
Doctrine	Historical data
Other Sources	Field test data
Analytic Rigor	Sensitivity Analysis
Comparison to valid models	Stress Test
Clarity and Economy	Black-box test
Relevant verisimilitude	Time-series Analysis
Experience/Intuition	Correlated Inspection
Existing Theory	Graph Analysis
Similar systems	Cause/Effect Graphing
Animation	Path Analysis
Walk-Through	Constraint Test
Formal Review	Inductive Assertions
Inspection	Boundary Analysis
Turing Tests	Traces
Event Validity	Extreme Condition Tests
Historical Methods	Fixed Values
	Predictive Validation
Peer Review	Internal Validity
	Historical Data Validation
	Degenerate Tests

Table A-1: Validation Techniques

A.1 Subjective Techniques

Face validation

Face validation is a rather bland term, and can be interpreted many ways. For this paper, face validation will consist of the model development team, along with system experts, formally discussing all the model's assumptions, the entities of the model, the variables of the model, the processes used and the output described in the model. This effort gives the system experts a chance to have an input into the model, during model development. If the system experts are convinced that the model is representing their system well, they can be a powerful ally in convincing management that the results are useful. Convincing the system experts that the model is a valid representation of the system under study also forces the model developer to examine his own work carefully. Face validation is basically a subjective comparison of the model and the real system. If there is no historical data to analyze, face validation can prove to be the most effective validation tool available. An example of such an instance would be the analysis of the feasibility of a new system and there is not an existing system to study. Dr. Gene Woolsey of the Colorado School of Mines goes even further in saying that to perform a valid simulation study, the analyst must actually be trained and work on the project before attempting to analyze the system.[1] Unfortunately, this is not a possibility for most analysts.

---

[1] Seminar, Dr. Gene Woolsey at the Air Force Institute of Technology, 1994.

Animation

Animation is a helpful validation tool, but can also be a pitfall. It can be very useful to get a visual confirmation that a model works as the analyst expected. It can also be a very good tool for selling the model results to management. If the manager can see a depiction of his system running on the computer, he is much more likely to trust the results. Animation can build more confidence in the manager than just delivering a list of numbers at the end of a report. Animation can be a hazard though. While animation can help in the understanding of the dynamic qualities of the model, it can lead to a false sense of security in believing the model is valid, just because it looks correct. A better idea is to use animation, if it is available, as a verification tool, and as a tool to prove that a model is not valid, instead of trying to prove that it is valid. (Law and Kelton, 1992; pg. 242) Animation should be used in a manner that it does not declare a model valid, rather it should be a test that the model must pass so that it is not declared invalid.

Walk-through

A Walk-through is similar to the inspection except that the team is concerned with standards and long-term implications. This effort adds to the overhead and does not appear to actually be a technique to increase validation confidence.

Formal reviews

These are structured efforts similar to inspections, but they are usually at a more general level of detail, and also involves management. Reviews should be scheduled

periodically, to keep management involved and to keep the analyst up to date on any large scale managerial changes that could affect the system and hence invalidate the conceptual model.

Inspection

An inspection is in fact a rather formal, structured, and large effort. It consists of a team of four or five analysts completing a formal list of steps to find faults. This includes 1) overview, 2) preparation, 3) inspection, 4) rework and 5) follow-up. This formal structure will probably make for a time consuming task.

Along the same lines is an effort called a Peer Review. None of the authors referenced in this work defines an effort exactly as a Peer Review. It is not a formal activity, like Balci's inspection, but rather the Peer Review is a face validation effort using a group of simulation analysts who are not associated with the project in question. Getting as much simulation experience together as possible and reviewing the conceptual model can be a large benefit to the validation effort.

Turing Tests

Turing tests consist of presenting two sets of data, one from the real system and one from model output, to system experts. The system experts then try to differentiate between the two sets, without prior knowledge as to which is which. This effort is presented by Balci (1995), Sargent (1994), and Law and Kelton (1991).

Event Validity

Events that occur in the simulation are compared to the real world system. The events do not have to be specific output of the model, rather an event can consist of any action that is performed by or on an entity in the model. Event validity is only possible if there is a method to track the events occurring during a simulation execution and if the events are comparable to the real-world occurrences. Simulation models contain abstractions that could make a one to one comparison with real-world events impossible.

Historical Methods

The historical methods of validating models are: Rationalism, Empiricism, and Positive Economics. Rationalism assumes that the underlying assumptions involved in a model are true, from these assumptions, logical deductions are made to create a valid model. Empiricism requires that all assumptions in the model must be validated experimentally. Positive Economics requires that the model only be able to predict the future correctly and is not concerned with the model's assumptions or structure used to achieve the results.

Peer Review

The peer review is the process of the validation analyst convincing other analysts who are not associated with the project that his conceptual model is good.

A.2 Empirical Techniques

Statistical analysis techniques

Statistical analysis covers a broad spectrum of topics. Use of statistical analysis techniques can be very powerful tools in validation, if used properly. The system needs to be observable (i.e., data can be collected) and the model has to be verified. Various methods can determine a confidence range for elements of the model, leading to overall validation of the model. The major stumbling block is the data. Very often, either the real world data is not in a usable format, if it exists. It is rare that an analyst can get the perfect data set needed. (Gass, 1991) Many times the system has no built in features to collect the data, or if it does, the level of detail of the data collected is so immense that the useful information cannot be sorted out. An example is log files of computer transactions. If the log file contains every action that the computer performed, sorting out the required information would be extremely difficult. When useful data is acquired, statistical analysis is one of the most powerful tools used. Statistical analysis techniques can yield objective, quantitative, reproducible data concerning the quality, or validity, of a simulation model. (Kleijnen, 1995)

Techniques such as Analysis of Variance, confidence intervals, Goodness of Fit tests, time series analysis, regression analysis, and F tests can be strong tools for validation. Several of the techniques are used as measures in hypothesis testing. All of these techniques are described in detail in Neter, Wasserman, and Kutner (1990).

Time series analysis is based on analysis of data relative to time. The output processes of most real-world systems and most simulations are not stationary processes (the distributions change over time) and are autocorrelated (Observations of the process are correlated with each other). (Law and Kelton, 1991) Under these conditions, classical statistical analysis based on independent, identically distributed (IID) observations cannot be directly used. However, there are many situations where time series analysis can be used. Details of time series analysis can be found in Neter, Wasserman, and Kutner (1990).

Basic concepts of time series analysis can be used regardless of any characteristics of the observations. Time series data and graphs can be analyzed for periodicity, max/min, inflection, skew, time periods of increase or decrease, etc. If they exist, basic time-series aspects are detectable and can be compared to known real world aspects.

Hypothesis testing is a concept used to test if the real world data and the model data could have come from the same population. Since the data sets could actually be looked at as samples from a population, hypothesis testing lets the analyst assign a degree of confidence to the nature of the relationship between the two data sets. There are several excellent software packages on the market for statistical analysis including SAS, Statistix, and Excel.

Sensitivity Analysis

Another important tool is sensitivity analysis. Varying inputs one at a time, and observing the changes in the output will identify variables that are important to the system

behavior. These variables can be used in a comparison to those in the real system. Most likely, this is used in the actual analysis of the system, via the model, after the model has been verified and validated. Sensitivity analysis is part of the information that the owner of the real system wants to find out about the performance of their system. If the proper data is available, sensitivity analysis can be used in validation. An example of it's possible use in validation is analysis where known phenomena exist. A model of a supersonic jet should have a marked change in performance when the speed of the plane is changed from mach .999 to mach 1. The changes in flight performance at the sound barrier are well documented and should be mimicked by the model, if the model was correct.

Very often, sensitivity analysis is done ad hoc. Normally, a few cases are used, where each variable is changed one at a time. While this can be useful, Kleijnen offers another approach that is a more scientific one, Response Surface Methodology (RSM). (Kleijnen, 1995) RSM uses polynomial response functions to approximate complex input/output relationships of a system. RSM consists of creating an experimental design of input variables that the analyst thinks might be important. Linear regression with first and second order terms is then used and the relationship between the input and output variables is approximated. The RSM design is then used for analysis in place of the actual relationship. Neter Wasserman, and Kutner (1990) has full details of carrying out the RSM analysis.

Stress Testing

Stress testing requires loading the model to it's maximum constraints and observing the model for any invalid response. Intuition on the analyst's part will be needed to discern if the behavior is an accurate prediction of real system behavior or if it is a problem with the model. If the model is showing poor results and there does not seem to be a good reason why, it could very well be indicating an error in the model. If no errors are detected from stress testing, the test will be included as part of the analysis of the system behavior.

Black-box testing

Black-box testing or functional testing, is an excellent method of validation, as long as all the proper historical data is available, and the model is verified to a high degree of confidence. Black-box testing consists of using test data in the model and checking if the resulting output is reasonably close to the actual output of the real system. This comparison requires statistical analysis, such as hypothesis testing, to compare the model outputs and the real-world system outputs. Most likely, an analyst will only be able to test a relatively small number of inputs. (Balci, 1994) Testing a limited range of inputs may lead the analyst to be suspicious of the actual validity over a larger range of inputs. The analyst should take careful consideration when choosing the inputs to cover as large a range as possible. For models with relatively small numbers of inputs and outputs, this task is manageable. Large, complex models can have millions of transformation paths between the inputs and outputs, which would be impossible to test. In this case, Response

Surface Methodology, which is described under sensitivity analysis, could be used in place of Black-box testing and achieve similar results.

Correlated inspection..

If the real system exists, hypothesis tests can be performed to determine if the simulation output data and the real-system data are distributed the same. Statistical analysis techniques as described under Balci are relevant for use. In addition, a technique called the Correlated Inspection approach can be useful. This technique compares the relative changes in the outputs from the simulation and the real-world system when using the same inputs. Comparing the relative changes, instead of the absolute results, attributed to identical inputs will show the correlation between the simulation and the real-world system. Since the model is an abstraction, it may not achieve the absolute results sought after, but may still achieve the correct relative results.

Graph based analysis

Graph based analysis is an exception that consists of creating flow charts of model control. These flow charts can help the analyst detect faults in the conceptual model during its creation. This technique seems to be more of a verification tool to check the implementation of the conceptual model.

Cause and Effect graphing

Cause and effect graphing can be used in conjunction with sensitivity analysis and addresses "what causes what". It is a graphical representation of which inputs and parameters affect output variables. This process requires analyzing the real system to determine the cause/effect relationships (possibly using the multivariate techniques described earlier) and then deciding if they are accurately described in the model. These relationships would be used in creating the model, if they were available. In this case, cause/effect graphing would become a verification effort. The cause/effect effort can become extremely large for large complex models. The HUNTOP model described by Kleijnen (1995) has over 40 input variables and a comparative number of outputs (unspecified). Creating a cause/effect graph for each of these would be a large, time consuming effort.

Path Analysis

Path analysis consists of testing all the control paths in the model. This analysis would be a good effort given unlimited time. Path analysis has the potential to become a very large task for a model of any complexity, since even small models will have many submodels. This type of analysis would be more of a verification tool, but could possibly identify validation errors. Testing a control path of a model is a difficult undertaking. The test requires executing data that will cause model control to pass into desired areas or submodels in the model. A software 'probe' would be required to track the flow of control

in the model. Limited path analysis is required in verification and debugging, but testing all control paths would be very time consuming for a model of any complexity.

Inductive Assertions

Inductive assertions (IA) is also mainly a verification tool, but could have some validation uses. IA consists of determining input-output relations, converting the relations into assertions, and checking the assertions at various points in the model's execution path. Checking the assertions, like several of the proceeding techniques, requires traceability along the execution path. The majority of errors detected would probably fall into verification, (Balci, 1994) but the test was included here because it is possible that it might detect validation errors also.

Boundary Analysis

Boundary analysis is similar to sensitivity analysis in that the analyst varies certain inputs by very small amounts to see the resulting changes in the output. The difference between boundary and sensitivity analysis is that the inputs varied are variables that have distinct regions, or domains, over their entire range. The variables are being tested at the boundaries of those regions. The reason why this test is included separately from sensitivity analysis is that the most error-prone cases lie near on the boarders of the variable ranges. (Balci, 1994)

Traces

Traces consist of following specific entities through the model and determine if the model's logic is correct.

Extreme-Condition Tests

The model should be tested for any conditions that maybe very unlikely, but are still possible. In such cases, the model should be bound if there are limitations on the actual operating ranges, such as a limited queue size for example. This test is similar, but not exactly, to Balci's stress test and constraint testing.

Fixed Values

All of the model's inputs and internal variables are set at fixed values to allow checking the model's results by calculation. This test sounds to be in the line of verification more than validation.

Predictive Validation

Predictive validation is determining if the model's prediction of the system behavior is accurate.

Internal Validity

Randomness of a stochastic model is checked by making multiple runs of the simulation model. A large amount of variability in the results may indicate a non-valid

model. If such a case occurs, the policy or system under consideration should be questioned.

Historical Data Validation

If enough historical data exists, the data is split into two groups. The first group is used to create the model, and the second group is used for validating the results of the model.

Degenerate Tests

The degeneracy (or state of becoming worse) of the model's behavior is tested by 1. Removing a section of the model, or 2. Making appropriate selections of values for inputs and parameters. An example of a degenerate test would be to increase the arrival rate to a queue until it is larger than the service rate and observe the resulting performance.

B. Case Study Techniques

Appendix B consists of details of particular validation techniques of interest in each of the case studies.

B.1 Case Study 1: RETACT

Table B-1 is a compilation of validation techniques reviewed in Appendix A. Details of the use of the techniques can be found in Appendix A. The techniques used in the RETACT analysis are highlighted with **Yes**.

Table B-1: RETACT Validation Techniques

Subjective (Informal)	Empirical (Formal)
Face Validation	Statistical Analysis
-Expert Opinion **Yes**	-Lab data **Yes**
-Doctrine	-Historical data **Yes**
-Other Sources	-Field test data **Yes**
-Analytic Rigor	Sensitivity Analysis
-Comparison to valid models	Stress Test
-Clarity and Economy	Black-box test
-Relevant verisimilitude	Time-series Analysis **Yes**
-Experience/Intuition **Yes**	Correlated Inspection
-Existing Theory	Graph Analysis
-Similar systems	Cause/Effect Graphing
Animation	Path Analysis
Walk-Through	Constraint Test
Formal Review	Inductive Assertions
Inspection	Proof of Correctness
Turing Tests	Traces
Event Validity	Extreme Condition Tests
Historical Methods	Fixed Values
	Predictive Validation
Peer Review	Internal Validity
	Historical Data Validation

B.2 Case Study 2: HUNTOP

Table B-2 is the compilation of validation techniques from Appendix A. Techniques that were used in the HUNTOP validation effort are designated by **Yes**. Details of the use of the techniques are in Appendix A.

Table B-2: HUNTOP Validation Techniques

Subjective (Informal)	Empirical (Formal)
Face Validation	Statistical Analysis
Expert Opinion **Yes**	Lab data
Doctrine	Historical data **Yes**
Other Sources	Field test data **Yes**
Analytic Rigor	Sensitivity Analysis **Yes**
Comparison to valid models	Stress Test
Clarity and Economy	Black-box test
Relevant verisimilitude	Time-series Analysis
Experience/Intuition	Correlated Inspection
Existing Theory **Yes**	Graph Analysis
Similar systems	Cause/Effect Graphing
Animation	Path Analysis
Walk-Through	Constraint Test
Formal Review	Inductive Assertions
Inspection	Proof of Correctness
Turing Tests	Traces
Event Validity	Extreme Condition Tests
Historical Methods	Fixed Values
	Predictive Validation
Peer Review	Internal Validity
	Historical Data Validation

B.3 Case Study 3: RADGUNS

Table B-3 is the compilation of validation techniques from Appendix A. Techniques that were used in the RADGUNS validation effort are designated by **Yes**. Details of the use of the techniques are in Appendix A.

Table B-3: RADGUNS Validation Techniques

Subjective (Informal)	Empirical (Formal)
Face Validation	Statistical Analysis
Expert Opinion **Yes**	Lab data **Yes**
Doctrine	Historical data **Yes**
Other Sources	Field test data **Yes**
Analytic Rigor	Sensitivity Analysis **Yes**
Comparison to valid models	Stress Test
Clarity and Economy	Black-box test
Relevant verisimilitude	Time-series Analysis **Yes**
Experience/Intuition	Correlated Inspection
Existing Theory **Yes**	Graph Analysis
Similar systems	Cause/Effect Graphing
Animation	Path Analysis
Walk-Through **Yes**	Constraint Test
Formal Review	Inductive Assertions
Inspection	Proof of Correctness
Turing Tests	Traces
Event Validity	Extreme Condition Tests
Historical Methods	Fixed Values
	Predictive Validation
Peer Review	Internal Validity
	Historical Data Validation

B.4 Star-Field Model

Table B-4 is the compilation of validation techniques from Appendix A. Techniques that were used in the Star-Field validation effort are designated by **Yes**. Details of the use of the techniques are in Appendix A.

Table B-4: Star-Field Validation Techniques

Subjective (Informal)	Empirical (Formal)
Face Validation	Statistical Analysis
Expert Opinion **Yes**	Lab data
Doctrine	Historical data **Yes**
Other Sources	Field test data **Yes**
Analytic Rigor	Sensitivity Analysis **Yes**
Comparison to valid models	Stress Test
Clarity and Economy	Black-box test
Relevant verisimilitude	Time-series Analysis
Experience/Intuition	Correlated Inspection
Existing Theory **Yes**	Graph Analysis
Similar systems	Cause/Effect Graphing
Animation	Path Analysis
Walk-Through	Constraint Test
Formal Review	Inductive Assertions
Inspection	Proof of Correctness
Turing Tests	Traces
Event Validity	Extreme Condition Tests
Historical Methods	Fixed Values
	Predictive Validation
Peer Review	Internal Validity
	Historical Data Validation

B.5 CERES-Wheat Model

Table B-5 is the compilation of validation techniques from Appendix A. Techniques that were used in the CERES-Wheat validation effort are designated by **Yes**. Details of the use of the techniques are in Appendix A.

Table B-5: CERES-Wheat Validation Techniques

Subjective (Informal)	Empirical (Formal)
Face Validation	Statistical Analysis
Expert Opinion	Lab data
Doctrine	Historical data
Other Sources	Field test data **Yes**
Analytic Rigor	Sensitivity Analysis **Yes**
Comparison to valid models	Stress Test
Clarity and Economy	Black-box test
Relevant verisimilitude	Time-series Analysis
Experience/Intuition	Correlated Inspection
Existing Theory	Graph Analysis
Similar systems	Cause/Effect Graphing
Animation	Path Analysis
Walk-Through	Constraint Test
Formal Review	Inductive Assertions
Inspection	Proof of Correctness
Turing Tests	Traces
Event Validity	Extreme Condition Tests
Historical Methods	Fixed Values
	Predictive Validation
Peer Review	Internal Validity
	Historical Data Validation

B.6 Fish Habitat Model

Table B-6 is the compilation of validation techniques from Appendix A. Techniques that were used in the Fish Habitat validation effort are designated by **<u>Yes</u>**. Details of the use of the techniques are in Appendix A.

Table B-6: Fish Habitat Validation Techniques

Subjective (Informal)	Empirical (Formal)
Face Validation	Statistical Analysis
Expert Opinion	Lab data
Doctrine	Historical data **Yes**
Other Sources	Field test data
Analytic Rigor	Sensitivity Analysis
Comparison to valid models	Stress Test
Clarity and Economy	Black-box test
Relevant verisimilitude	Time-series Analysis
Experience/Intuition	Correlated Inspection
Existing Theory	Graph Analysis
Similar systems	Cause/Effect Graphing
Animation	Path Analysis
Walk-Through	Constraint Test
Formal Review	Inductive Assertions
Inspection	Proof of Correctness
Turing Tests	Traces
Event Validity	Extreme Condition Tests
Historical Methods	Fixed Values
	Predictive Validation
Peer Review	Internal Validity
	Historical Data Validation

BIBLIOGRAPHY

1. U.S. Air Force Directorate of Modeling, Simulation and Analysis (AF/XOM). Air Force Instruction 16-1001. Washington DC 1995.

2. Army Regulation 5-11. Army Model and Simulation Management Program. HQ, Department of the Army. Washington DC, 1992.

3. Arquilla, John and Davis, Paul. Modeling Decision making of Potential Proliferators as Part of Developing Counterproliferation Strategies. RAND Corp. Santa Monica, CA, 1994.

4. Bacsi, Z., and Zemankovics, F., "Validation: and objective or a tool? Results on a winter wheat simulation model application." Ecological Modeling, vol. 81, August, 1995.

5. Balci, Osman, Validation, Verification, and Testing Techniques throughout the Life Cycle of a Simulation Study. Virginia Polytechnic Institute, 1994.

6. Balci, Osman, "Credibility assessment of simulation results: The state of the art" Methodology and Validation, Simulation Councils Inc., San Diego Ca., 1987.

7. Davis, Paul. Generalizing Concepts and Methods of Verification, Validation and Accreditation for Military Simulations. RAND Corp. Santa Monica Ca., 1992.

8. Department of Defense Directive Number 5000.59. DoD Modeling and Simulation Management, 1994.

9. Gass, S.I., Fossett, C., Harrison, D., and Weintrob, H. "An Assessment Procedure for Simulation Models: A Case Study", Operations Research, Vol. 39, No. 5, 1991.

10. Gass, S I. and Joel, L. S. "Concepts of Model Confidence" Computer Operations Research Vol. 8, no. 4, 1981.

11. Hodges, James and Dewar, James. Is it You or Your Model Talking? A Framework for Model Validation. RAND Corp. Santa Monica, CA, 1992.

12. Kleijnen, Jack. Statistical Validation of Simulation Models. Tilburg University, Tilburg, Netherlands. 1995.

13. Kleindorfer, George and Ganeshan, Ram. The Philosophy of Science and Validation in Simulation. Pennsylvania State University, University Park, PA. 1993.

14. Law, Averil and Kelton, W. David. Simulation Modeling and Analysis. McGraw-Hill, New York, NY., 1991.

15. Landry, M. And Oral, M. "In Search of a valid view of model validation for operations research." European Journal of Operational Research Vol. 66, no 2, 1993.

16. Mayer, Hugo. Approximations, Assumptions and Validation. TRADOC Analysis Center, Fort Leavenworth, Ka., 1995.

17. McVoy, C.W., Kersebaum, K., Arning, M., Kleeberg, P., Othmer, H., and Schroeder, U., "A Data Set from North Germany for the Validation of Agroecosystem Model: Documentation and Evaluation." Ecological Modeling vol. 81, August, 1995.

18. Naylor, T. H. And Finger, J. M. "Verification of Computer Simulation Models." Management Science, No. 14, 1967.

19. Neter, John, and Wasserman, William, and Kutner, Michael. Applied Linear Statistical Models. Irwin. Burr Ridge, Ill. 1990.

20. Nicholas, John M., Managing Business and Engineering Projects: Concepts and Implementations. Englewood Cliffs, New Jersey. Prentice Hall, Inc. 1990.

21. Sargent, Robert, Verification and Validation of Simulation Models. Proceedings of the 1994 Winter Simulation Conference, 1994.

22. Sargent, Robert, "Some Comments on Model Validation" Computer Modeling and Simulation: Principles of Good Practice. Simulation Councils Inc., La Jolla, Ca., 1982.

23. Schlesinger, S., Buyan, J., Callender, E., Clarkson, W., Perkins, F. "Developing Standard Procedures for Simulation Validation and Verification". 1974 Summer Computer Conference Proceedings. 1974.

24. Shannon, Robert, Systems Simulation: The Art and Science Prentice-Hall, Englewood Cliffs, NJ. 1975.

25. Stefan, H., Hondzo, M., Eaton, J., McCormick, J., "Validation of a fish habitat for lakes" Ecological Modeling, vol. 82, August 1995.

26. Susceptibility Model Assessment and Range Test Project (SMART). An Accreditation Support Framework for DoD Models and Simulations. Draft Version. Naval Air Warfare Center, China Lake, Ca., 1995.

27. Susceptibility Model Assessment and Range Test Project (SMART). Comparative Analysis of Tri-service Accreditation Policies and Practices. Draft Version. Naval Air Warfare Center, China Lake, Ca., 1995.

28. Susceptibility Model Assessment and Range Test Project (SMART). Information Requirements in Support of Accreditation. Draft Version. Naval Air Warfare Center, China Lake, Ca., 1995.

29. Susceptibility Model Assessment and Range Test Project (SMART). Accreditation Support Package for RADGUNS. Draft version.Naval Air Warfare Center, China Lake, Ca., 1995.

30. Piplani, Lalit, Colonel, US Army, System's Acquisition Manager's Guide for the use of Models and Simulations. Defense Systems Management College Press, Fort Belvoir, Va., 1994.

31. Wedberg, George. "But First, Understand the Problem" Journal of Systems Management. Pg. 20-28, June, 1990.

32. Winter, E. and Wisemiller, D., Validation of Simulated Sensor Observations of Star Fields. Proceedings of the 1974 Summer Computer Simulation Conference, La Jolla Ca. 1974.

Vita

Captain Michael Elmer was born on December 7, 1967, in Akron Ohio to Otto and Marie Elmer. He is the seventh of seven children in the Elmer family. He graduated Firestone High School in 1986 and continued on to Case Western Reserve University in Cleveland Ohio. While at Case Western, he became a member of the Fraternity of Phi Gamma Delta and discovered his one true passion, bicycling. After graduating from Case Western in 1990 with a bachelor's degree in Systems Engineering, he was commissioned a second lieutenant in the United States Air Force and was assigned to the Standard Systems Center at Gunter AFB in Montgomery Alabama. He served at Gunter for just under three years before being reassigned to the Air Force Institute of Technology at Wright Patterson AFB in Dayton Ohio. Upon graduation from AFIT with a master's degree in Space Operations, he is scheduled to be assigned to U.S. Space Command at Peterson AFB in Colorado Springs Colorado where he plans to ride his bicycle up Pikes Peak.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 1996	Master's Thesis

4. TITLE AND SUBTITLE
Issues and Challenges in Validating Military Simulation Models

5. FUNDING NUMBERS

6. AUTHOR(S)
Michael R. Elmer, Capt, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology, WPAFB OH 45433-6583

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Ray Miller III
AF/XOMT
1480 Air Force Pentagon
Washington, DC 20330-1480

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The purpose of this thesis is to address the challenges of validating simulation models, especially those challenges facing military simulation analysts. Three distinct issues are of concern to the military simulation analyst: 1) What type of validation effort do the academic experts recommend? 2) What does the military policy say is necessary for a proper validation effort? 3) What can a simulation practitioner realistically accomplish given time and resource constraints? Four methodologies were chosen to represent the academic perspective on validation. A model of validation methods is integrated from the methodologies of these four simulation validation references. The validation policies of the Army, Navy, and Air Force are examined and analyzed for their methodologies to be applied to simulations. The integrated model is compared to these policies that are being formed within the DoD to determine the relationship between the academic experts and the military policies. Case studies of validation efforts are examined and analyzed for the methodologies used by simulation practitioners. The integrated model is compared to the case studies to examine the relationship between the academic experts and the actual practitioners. Finally, conclusions and observations are drawn from all of these comparisons.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Validation, Simulation Model, Military	107
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL